New Computational Methods for the Investigation of Thermodynamics and Kinetics of Protein Folding

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Stefanie Muff

aus

Neuenkirch LU

Promotionskomitee

Prof. Dr. Amedeo Caflisch (Vorsitz)
Prof. Dr. Benjamin Schuler

Zürich 2008

Summary

There is a large demand for methods to analyze fully atomistic molecular dynamics (MD) trajectories, which is very challenging due to the many degrees of freedom involved. New methods for extracting thermodynamics and kinetics from MD simulations of proteins are presented in this thesis. The ultimate goal of such methods is to fully quantify the free-energy surfaces of peptides and small proteins, which can nowadays be simulated at folding-unfolding equilibrium or with the replica exchange molecular dynamics (REMD) method.

Two related procedures to address this problem are introduced and discussed here. The main idea behind both is that conformations are grouped not according to a structural similarity criterion, but according to the transitions observed in equilibrium simulations, i.e., if conformations are kinetically similar. The *kinetic grouping analysis* (KGA) requires a parameter, the commitment time τ_{commit}, which is a typical relaxation time within the free-energy basins of the system. On the other hand, the *cut-based free-energy profile* (cFEP) approach is parameter-free and requires only the selection of a representative conformation in a region of interest. Both KGA and cFEPs were successful in quantifying the free-energy basins of a β-sheet and a helical peptide. The analytical power of the cFEP method goes beyond the one of KGA because with the former it is possible to determine the correct barrier to leave any free-energy basin of interest. This important aspect of the cFEP method can be employed to isolate the unfolding transition state region, which corresponds to the ensemble of conformations on top of the unfolding barrier.

In the present thesis, the cFEP method was applied not only to the analysis of equilibrium dynamics, but also to data obtained by REMD simulations. The latter is an enhanced sampling technique which is able to recover correct thermodynamics (i.e., population of states), but not the kinetics. This problem was addressed here by supplementing the cFEPs at a constant temperature with kinetic information from higher temperatures by scaling the folding times according to the Arrhenius equation.

Zusammenfassung

Die vorliegende Doktorarbeit befasst sich mit neuen Methoden zur Bestimmung von thermodynamischen und kinetischen Eigenschaften von Proteinen, welche mit Moleküldynamik im Computer simuliert werden. Die Herausforderung bei der Entwicklung solcher Methoden besteht darin, dass durch die explizite Darstellung aller Atome eines Proteinmoleküls in der Simulation die Anzahl Freiheitsgrade drastisch ansteigt. Das Ziel ist die präzise quantitative Beschreibung der Freien-Energie-Oberfläche eines solchen Moleküls. Dies beinhaltet die Bestimmung aller metastabilen Zustände, Überganszustände und der Geschwindigkeit, mit der sich das System zwischen den Zuständen bewegt. In dieser Arbeit werden Ansätze zur Analyse von Daten aus Simulationen im thermischen Gleichgewicht oder aus Replica Exchange Molecular Dynamics (REMD) Simulationen diskutiert. Letztere ist eine Simulationstechnik, welche die statistischen Eigenschaften des simulierten Konformationsraumes verbessert.

Die Kernidee in den beiden hier besprochenen Verfahren ist, dass Konformationen nicht aufgrund struktureller Eigenschaften in Freie-Energie-Minima gruppiert werden, sondern nach ihrer kinetischen Ähnlichkeit, d.h. wenn sie schnell ineinander übergehen können. Die *kinetic grouping analysis* (KGA) stützt sich dabei auf einen Parameter τ_{commit}, welcher die typische Relaxationszeit innerhalb der Minima des Systems repräsentiert. Die zweite Methode der *cut-based free-energy profiles* (cFEPs) hingegen braucht keine Parameter, sondern lediglich die Wahl eines Repräsentanten der Region, die man untersucht. Beide Verfahren wurden erfolgreich zur quantitativen Bestimmung der Freien-Energie-Minima von Peptiden mit einer β-Faltblatt- und einer Helixstruktur eingesetzt. Der Vorteil der cFEP-Methode ist, dass sie auch zur Bestimmung der Höhe von Energiebarrieren eingesetzt werden kann und sehr genau bestimmt, wo die Faltungs- und Entfaltungs-Übergangszustände liegen.

Das cFEP-Verfahren konnte nicht nur erfolgreich zur Analyse von Gleichgewichts-Simulationen eingesetzt werden, sondern half auch, REMD-Daten zu quantifizieren. REMD liefert ein thermodynamisch korrektes Ensemble, dafür geht die kinetische Information verloren. Durch die Skalierung von Faltungsraten bei höheren Temperaturen unter der Verwendung der Arrheniusgleichung konnte ein Teil der kinetischen Information zurückgewonnen werden.

List of publications

Local modularity measure for network clusterizations.
S. Muff, F. Rao, A. Caflisch
[*Phys. Rev. E*, **2005**, *72*, 056107]

Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein.
S. Muff and A. Caflisch.
[*Proteins: Structure, Function and Bioinformatics*, **2008**, *70(4)*, 1185-1195]

α-helix folding in the presence of structural constraints.
J.A. Ihalainen, B. Paoli, S. Muff, E.H. Backus, J. Bredenbeck, G.A. Woolley, A. Caflisch, P. Hamm
[*PNAS*, **2008**, *105(28)*, 9588-9593]

One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: new insights into the folding process.
S.V. Krivov[‡], S. Muff[‡], A. Caflisch, M. Karplus
[‡] These two authors contributed equally to this work
[*J. Phys. Chem. B*, **2008**, *112(29)*, 8701-8714]

Identification of the protein folding transition state from molecular dynamics trajectories.
S. Muff and A. Caflisch
[*Submitted*]

ETNA: Equilibrium transition networks and Arrhenius equation for extracting folding kinetics from REMD simulations.
S. Muff and A. Caflisch
[*Submitted*]

Contents

Chapter 1

Towards the understanding of the protein folding free-energy surface

1.1 Introduction

Proteins are complex macromolecules involved in a large variety of physiological processes in all organisms. Most proteins fulfil their function if they are folded to a unique native structure, also termed as the folded state. A protein, however, can attain a very large number of non-native three-dimensional structures, and the folding process is a complex reaction because a system of N atoms has 3N degrees of freedom [1].

The folding process is governed by two main driving forces. On one hand, enthalpic stabilization E depends on van der Waals and electrostatic interactions among protein atoms, as well as on the effect of the solvent. On the other hand, the entropy S, which accounts for the flexibility of the protein, is important because an unfavorable enthalpic energy can be counterbalanced by a large number of possible states at that energy. Therefore, the movement of the molecule depends on two competing contributions, which result in the free energy G [2]. At temperature T the free energy of a state i is written as

$$G_i = E_i - TS_i \ .$$

As mentioned above, both enthalpy and entropy play a major role in protein folding. The understanding of the free-energy surface is therefore crucial, i.e., analysis of only the potential-energy surface is insufficient [3–6]. There are major aspects of interest in protein folding that require the investigation

of the free-energy surface. If the full knowledge of the surface is available, the following quantities can be determined:

(i) population and structure of free-energy *basins*, associated with local free-energy minima.

(ii) the stability of basins, reflected by the *barriers* separating free-energy minima from each other.

(iii) the *diffusion coefficient* in any point of the surface.

(iv) folding and unfolding *rates*, calculated directly from the knowledge of all barriers and the diffusion coefficient.

(v) the *transition state ensemble* (TSE), which corresponds to the top of the unfolding barrier.

In the present thesis, the protein folding free-energy surface of the designed peptide Beta3s was studied (Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17} -Ile_{18}-Tyr_{19} -Thr_{20}). Beta3s folds in solution to the three-stranded antiparallel β-sheet shown in Figure 1.1 [7, 8]. All studies were performed *in silico*, i.e, by running computer simulations of molecular dynamics (MD). The fact that it was possible to perform long folding-unfolding equilibrium MD trajectories, at least with an implicit solvation model [9], makes the system well suited as a model for methodological developments to describe free-energy surfaces. A 20 μs trajectory at 330 K was the basis of the analyses in the following sections.

In the rest of Chapter 1 a short overview over the history of different methods to capture the features of free-energy surfaces is given. The first approach, which is still widely used, is the projection of the complexity onto order parameters of one or at most a few dimensions. The inadequacy of arbitrary projections is pointed out and three methods that address this problem are discussed.

1.2 The projection onto order parameters

The information harvested from *in vitro* experiments is usually limited by a small number of observables, meaning that only very few dimensions of the system can be captured. On the other hand, MD simulations contain all geometrical and dynamical details, and the challenge is to deal with this overwhelming amount of information. A common way to reduce the complexity is by projecting onto one or a few observables, the *order parameter(s)* [10], which are then used to describe the free-energy surface. Both, low-dimensional experimental signals and projections onto order parameters may lead to an oversimplified picture of folding, as will be discussed later in this chapter and exemplified for the Beta3s system. The latter has been

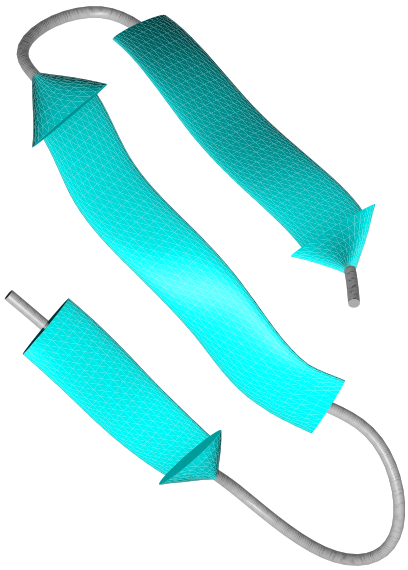


Figure 1.1: **Native state conformation of Beta3s.** The solution conformation has been studied by NMR [7] and the data indicate that Beta3s is a monomeric, three-stranded antiparallel β-sheet. The turns are formed by Gly_6-Ser_7 and Gly_{14}-Ser_{15}.

analyzed by the projection onto the fraction of native contacts Q [8,11]. The free energy $G(Q)$ was deduced from the Boltzmann distribution as

$$G(Q) = -kT \log(P(Q)) \ ,$$

where $P(Q)$ is the probability that the system has formed the fraction Q of the 26 possible native contacts, k is the Boltzmann constant and T is the temperature. The emerging Beta3s free-energy surface suggests a simple two-state picture of folding (Figure 1.2).

However, it was shown that the denatured state ($Q \approx 0.3$) is very heterogeneous [12, 13]. Furthermore, the putative transition state ensemble (TSE; $Q \approx 0.6$) contains not only structures with a folding probability $p_{fold} \approx 0.5$, but also many with $p_{fold} \approx 0$ [13], which are obviously not belonging to the TSE. Generally, the problem with order parameters such as Q is that they are chosen arbitrarily and can hide important aspects of the folding process [14]. Finding informative order parameters for folding is very difficult [15, 16] and is successful only for simple systems, such as the alanine dipeptide, which can be described entirely by a few dimensions [16–18].

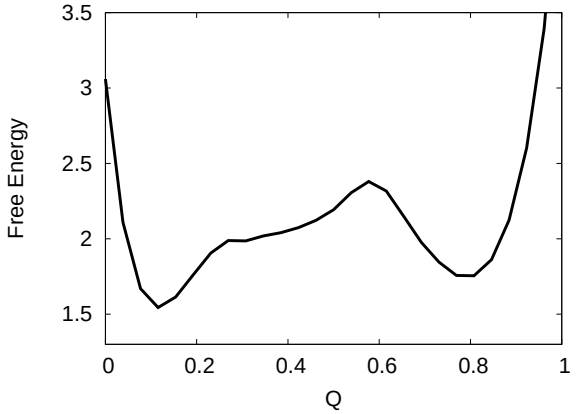


Figure 1.2: **Reduction of complexity.** In the case of Beta3s, the projection onto the fraction of native contacts Q erroneously suggests that the system is a two-state folder.

1.3 The protein folding network

To overcome the limitations of projected free-energy surfaces, an approach based on the theory of complex networks [19] was developed in our group [12, 20, 21]. The result is a *qualitative* picture of the multi-dimensional free-energy surface, whose complexity is preserved. The network analysis is based on a discretization of the conformational space, e.g., rmsd or secondary structure coarse-graining [12, 14, 22, 23]. Snapshots with similar structural properties are grouped together into *nodes* (also called conformations hereafter), while transitions from one node to another during the MD trajectory establish the *links*.

Figure 1.3 shows the protein folding network of Beta3s. The network reveals a very heterogeneous unfolded state with multiple metastable regions. Within the network framework, these "communities" correspond to the free-energy basins of the system. Their isolation is equivalent to finding a "good" clusterization of the network [21,24] (Chapter 7). Closer inspection of the basins indicates that some are stabilized entropically, others enthalpically [12]. However, the analysis of the basins is qualitative in the sense that no populations or barriers can be estimated from the picture. This problem is engraved because the network contains only nodes that are populated by a significant number of snapshots (40 in the Beta3s networks), in order to prevent overcrowding. From all snapshots, less than 55% are represented in the picture of Figure 1.3, because a large number of nodes are populated by only very few snapshots, most of them by only one or two. Therefore, even though the network reveals the presence of multiple (enthalpic) basins, it

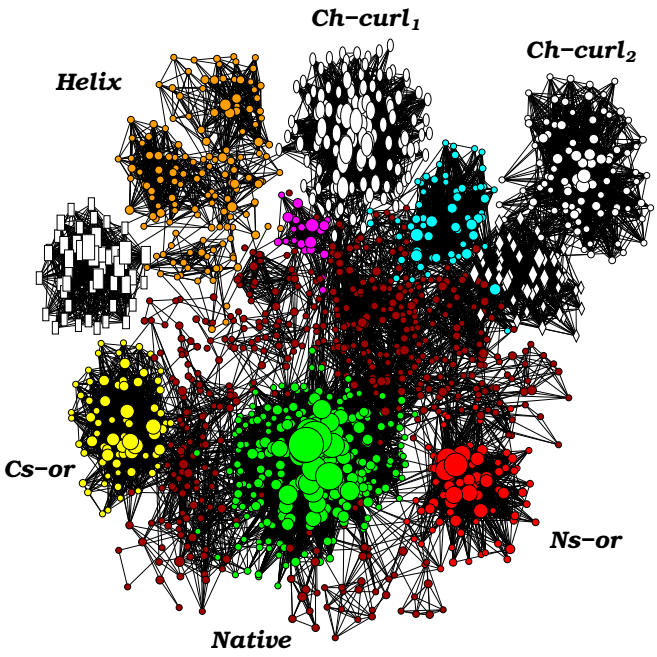


Figure 1.3: **Network representation of a 20 μs simulation of Beta3s.** Nodes populated by at least 40 snapshots are represented in the figure. Colors indicate different basins, as isolated with the kinetic grouping analysis.

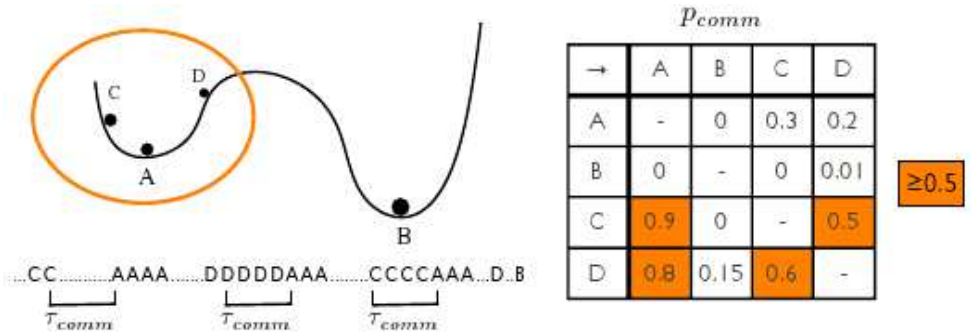
might miss some other aspects of the free-energy surface, especially in the presence of entropic regions with many low populated nodes.

1.4 Kinetic grouping analysis (KGA)

One important aspect of the analysis of free-energy surfaces is the quantitative description of free-energy basins [25,26]. Often, such analysis is based on structural similarity [23], meaning that snapshots fulfilling a certain structural criterion are grouped into a basin. However, similar structures can be separated by considerable barriers, for instance if the transition between them implies the rearrangement of sidechains [26]. Therefore, structural similarity does not guarantee the absence of barriers between two conformations.

On the other hand, if two conformations belong to the same basin,i.e., are not separated by a barrier, they are able to interconvert rapidly during the simulation [14]. An approach based on exactly this observation is called *kinetic grouping analysis* (KGA) [26]. KGA groups conformations according to fast relaxation within equilibrium trajectories and requires only one parameter, the commitment time τ_{commit}, which is a typical relaxation

time within the basins of the system. Figure 1.4 illustrates these ideas for a simple two-state system, which is discretized into 4 nodes. Obviously, A, C and D belong to the same basin, B is separated by a barrier. The table on the right of Figure 1.4 contains the commitment probabilities p_{commit}, i.e., the probabilities that one node interconverts to another within τ_{commit} during the trajectory. If the commitment time is chosen appropriately [26], then nodes which interconvert with $p_{commit} \geq 0.5$ belong to the same basin. Nodes C and D relax to A (the bottom of the left basin) with high p_{commit} and are therefore grouped together. Interconversion from and to node B is much slower, which is reflected in very low p_{commit} values between B and all other nodes.



p_{comm}

→	A	B	C	D
A	-	0	0.3	0.2
B	0	-	0	0.01
C	0.9	0	-	0.5
D	0.8	0.15	0.6	-

≥0.5

Figure 1.4: **Illustration of the KGA procedure (Chapter 2).** Nodes A, C and D belong to the same basin and therefore interconvert rapidly within the timeseries. As a result, p_{commit} values of C and D to the bottom node A are ≥ 0.5 and KGA groups them together. B stays alone, since the barrier prevents fast relaxation between B and the other nodes.

The procedure illustrated above is a *quantitative* method that determines the relative population of basins. KGA can be applied straightforwardly to very complex systems, like Beta3s [26] (Chapter 2) or a cross-linked α-helical peptide [27] (Chapter 3). In the case of the latter, the relative population of basins was evaluated with a τ_{commit} of 1 ns and the result is reflected qualitatively in the colorization of Figure 1.3. By applying the same analysis to a 20 μs simulation of the W10V mutant of Beta3s, it was possible to quantify changes in the free-energy surface upon a small mutation [26].

Note, however, that the KGA method is suitable to isolate mainly enthalpically stabilized basins, because the assumption of fast relaxation does not hold in very entropic regions, where most nodes are visited only once or a few times. Another caveat of the KGA method is that barriers and rates cannot be isolated directly.

1.5 Cut-based free-energy profiles

Like the KGA introduced above, the grouping of conformations according to equilibrium dynamics instead of geometrical characteristics is the essential aspect of transition disconnectivity graphs (TRDGs) [14, 28]. To this aim, the equilibrium transitions network (ETN) is constructed as described in Section 1.3. The main assumption is that the ETN contains the same kinetic and thermodynamic information as the MD trajectory from which it was constructed, but in a more condensed and informative form. Barriers between any two nodes of interest are determined by calculating the minimum-cut through the network that separates the two nodes, which – according to the Ford-Fulkerson theorem – corresponds to the "maximal flow", i.e., the barrier between the nodes, if all routes through the network are taken into account [14]. It is possible to isolate all basins and barriers by iterative determination of minimum-cuts between all pairs of nodes with the Gomory-Hu algorithm [29], and this is how the TRDG is calculated.

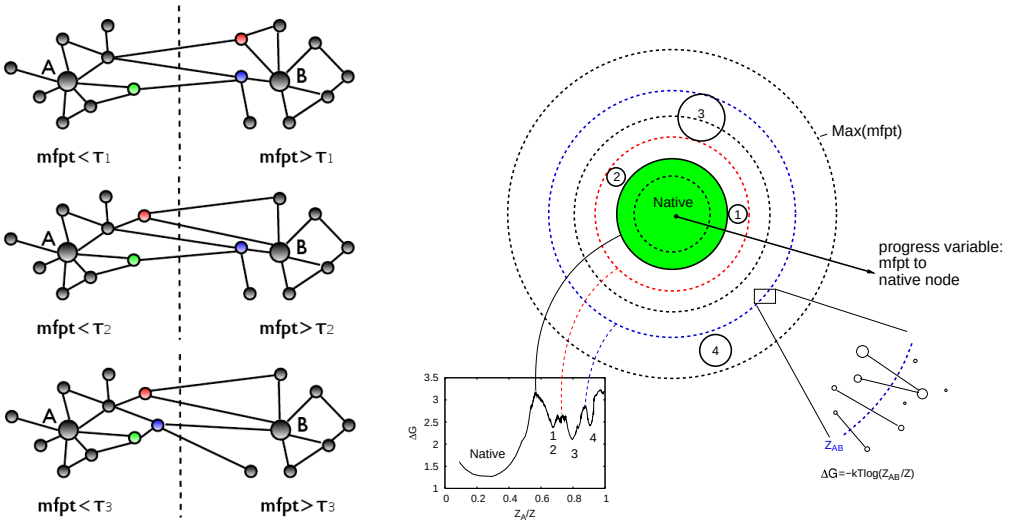


Figure 1.5: **Illustration of the procedure used to calculate the mfpt-cFEP (Chapter 4).**

Later, the minimum-cut procedure has been reinterpreted by Krivov and Karplus and was used for the calculation of one-dimensional, barrier-preserving profiles, so-called *cut-based free-energy profiles* (cFEPs), where the relative partition function was used as a progress coordinate [25]. However, the calculation of minimum-cuts is not only expensive, but also limited to a subset of points. Therefore, cFEP methods based on the progress variables p_{fold} [25] and the mean first passage time (mfpt) [30] were introduced and shown to approximate the exact barrier very well.

Figure 1.5 illustrates the cFEP procedure using mfpt as the progress variable. First, the mfpt to the most populated (i.e., native) node is calculated analytically on the ETN (with boundary condition mfpt=0 for the native node), and all nodes are sorted according to increasing values of mfpt. Then, for each value mfpt_c the relative partition function Z_A/Z is the fraction of all nodes with mfpt<mfpt_c (x-axis). Z_{AB}/Z is the cut, i.e., the normalized sum of all links connecting nodes with mftp<mfpt_c and mfpt>mfpt_c, from which the the free-energy barrier between these two sets of nodes can be calculated as $\Delta G = -kT \log(Z_{AB}/Z)$ (y-axis). Remarkably, the procedure requires no parameters.

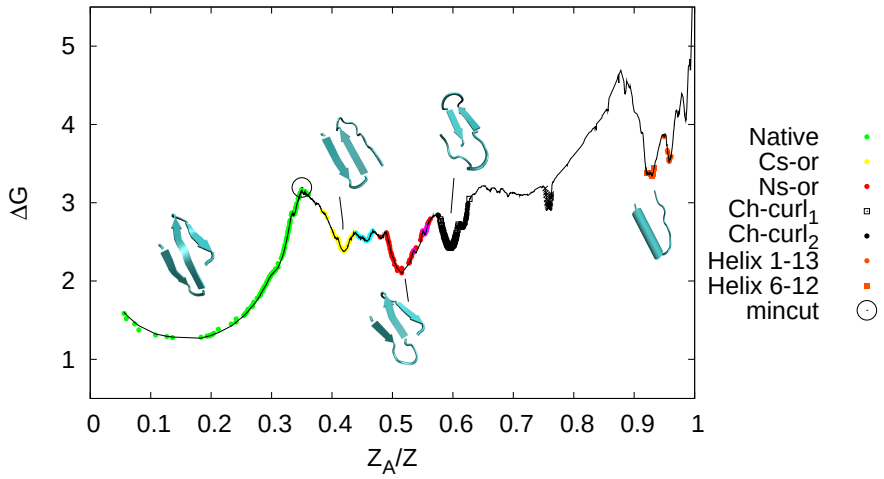


Figure 1.6: **cFEP of Beta3s.** Symbols represent nodes populated by ≥ 100 snapshots; their color code is the same as in Figure 1.3, i.e., it corresponds to the basins identified by the KGA. The native basin (green) populates the region with Z_A/Z <0.35. Exact populations of other basins can be determined by plotting separate profiles [30].

Figure 1.6 shows the unfolding cFEP of Beta3s, where the most populated (native) node has boundary condition mfpt=0. All nodes lying on the left of the cut at the first barrier belong to the native basin (Z_A/Z <0.35, i.e., a population of 35%) and the first barrier corresponds to the unfolding barrier. As mentioned above, the cFEP barrier is a very good approximation to the exact value, which can be calculated from the minimum-cut (black circle). In order to systematically identify all basins and barriers, the respective profiles from nodes in other regions have been plotted [25, 30]. The cut at the first barrier determines the basin of the reference node and the barrier represents the activation free-energy to exit. Interestingly, it was possible to identify a helical and a heterogeneous entropic region populated by about

11.6% and 33%, respectively, which KGA failed to reveal. All other basins isolated from cFEPs were identical to those from KGA and therefore the two approaches validate each other [30].

Other systems, such as the β-hairpin of protein G [25] and a lattice protein model [31] were also successfully analyzed with the cFEP approach. In the presence of equilibrium simulation data, the method is able to fully quantify the free-energy surface. Free-energy basins and barriers can be exactly determined, which in turn makes possible the extraction of kinetics.

1.6 Isolation of the transition state

The TSE is the ensemble of structures at the critical point in the folding process, where the molecule is equally probable to fold or to unfold. Note that structures in the folding TSE have $p_{fold} \approx 0.5$, but, depending on how p_{fold} is defined, not all structures with $p_{fold} \approx 0.5$ do necessarily belong to the TSE [25]. The pictorial interpretation in terms of free-energy surfaces is to describe the TSE as the ensemble of structures lying on top of the (un)folding barrier. Many approaches have been proposed to identify putative TSE structures from MD trajectories [13, 32–35] and a method based on p_{fold} [10] has been used for validating them [34, 36–38]. It was claimed in the previous section that the cFEP approach is able to entirely describe the free-energy surface. If this holds, then the successful identification of the (un)folding TSE from the profiles is a consequence.

As a proof of concept, shooting simulations from a selection of nodes along the Beta3s cFEP were performed. These simulations are short MD runs, started from structures in the selected nodes many times and with different initial velocities. One run is considered to be a successful folding event, if the trajectory visits the native node within a certain commitment time τ_{commit}, which has to be chosen much shorter than the folding time, but long enough to allow local relaxation [36]. The fraction of successive folding events among all runs started from a node is its p_{fold}.

The black squares in Figure 1.7 mark the nodes chosen for shooting simulations. The p_{fold}-values were evaluated by starting a total of 200 runs for each node. The results confirm that the top of the first free-energy barrier in the cFEP corresponds to the $p_{fold} \approx 0.5$ region, i.e., to the TSE of the system. Nodes before the first barrier belong to the native basin and are correctly assigned a $p_{fold} \approx 1$, while nodes after the barrier have $p_{fold} \approx 0$. These results indicate that the cFEP approach is able to correctly identify not only basins and barriers, but also the TSE to exit the region of interest (usually the folded state). Since no parameter is required in all the analysis, the cFEP method is an objective and effective method that will probably find many applications in the future.

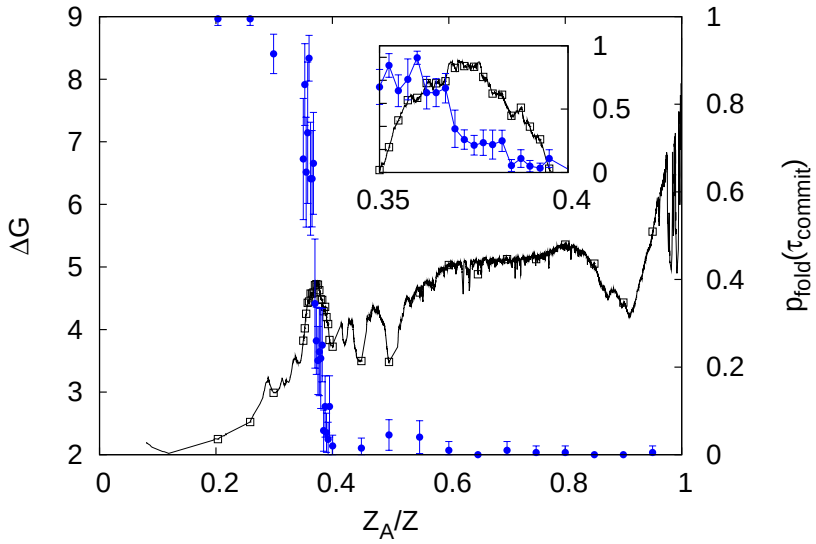


Figure 1.7: **Folding TSE identification from the Beta3s cFEP (Chapter 5).** p_{fold} values (blue), as evaluated from shooting simulations for a selection of nodes chosen along the profile (black squares), confirm that the top of the first barrier corresponds to the folding TSE, which has $p_{fold} \approx 0.5$.

Bibliography

[1] M. Karplus. Aspects of protein reaction dynamics: deviations from simple behavior. *J. Phys. Chem. B*, 104:11–27, 2000.

[2] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscape and motions of proteins. *Science*, 254:1598–1603, 1991.

[3] A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in Biochemical Sciences*, 25:331–339, 2000.

[4] L. A. Mirny and E. I. Shakhnovich. Protein folding theory: from lattice to all-atom models. *Ann. Rev. Biophys. Biomolec. Struc.*, 30:361–396, 2001.

[5] V. Daggett and A. R. Fersht. The present view of the mechanism of protein folding. *Nature Rev. Mol. Cell Biol.*, 4:497–502, 2003.

[6] P. G. Wolynes. Energy landscapes and solved protein-folding problems. *Phil. Trans. R. Soc. A*, 363:453–467, 2005.

[7] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez. De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Science*, 8:854–865, 1999.

[8] P. Ferrara and A. Caflisch. Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci. USA.*, 97:10780–10785, 2000.

[9] P. Ferrara, J. Apostolakis, and A. Caflisch. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 46:24–33, 2002.

[10] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108:334–350, 1998.

[11] A. Cavalli, P. Ferrara, and A. Caflisch. Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Structure, Function, and Bioinformatics*, 47:305–314, 2002.

[12] F. Rao and A. Caflisch. The protein folding network. *J. Mol. Biol.*, 342:299–306, 2004.

[13] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch. Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.*, 122:184901, 2005.

[14] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA.*, 101:14766–14770, 2004.

[15] V. S. Pande, A. Y. Grosberg, T. Tanaka, and D. S. Rokhsar. Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.*, 8:68–79, 1998.

[16] P. G. Bolhuis, C. Dellago, and D. Chandler. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. USA.*, 97:5877–5882, 2000.

[17] J. Apostolakis, P. Ferrara, and A. Caflisch. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J. Chem. Phys.*, 110:2099–2108, 1999.

[18] A. Ma and A. R. Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109:6769–6779, 2005.

[19] M. E. J. Newman. The structure and function of complex networks. *SIAM REV.*, 45:167–256, 2003.

[20] A. Caflisch. Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.*, 16:71–78, 2006.

[21] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. USA.*, 104:1817–1822, 2007.

[22] J.A. Hartigan. Clustering algorithms. *Wiley, New York*, 1975.

[23] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich. Understanding ensemble protein folding at atomic detail. *Proc. Natl. Acad. Sci. USA.*, 103:17747–17752, 2006.

[24] S. Muff, F. Rao, and A. Caflisch. Local modularity measure for network clusterizations. *Phys. Rev. E*, 72:056107, 2005.

[25] S. V. Krivov and M. Karplus. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B*, 110:12689–12698, 2006.

[26] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics*, 70:1185–1195, 2008.

[27] J. A. Ihalainen, B. Paoli, S. Muff, E. Backus, J. Bredenbeck, G. A. Woolley, A. Caflisch, and P. Hamm. α-helix folding in the presence of structural constraints. *Proc. Natl. Acad. Sci. USA.*, 105:9588–9593, 2008.

[28] S. V. Krivov and M. Karplus. Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys.*, 117:10894–10903, 2002.

[29] R.E. Gomory and T.C. Hu. Multi-terminal network flows. *SIAM J. Appl. Math.*, 9:551–570, 1961.

[30] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-dimensional barrier preserving free-energy projections of a beta-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B*, 2008. in press.

[31] A. Y. Palyanov, S. V. Krivov, M. Karplus, and S. F. Chekmarev. A lattice protein with an amyloidogenic latent state: Stability and folding kinetics. *J. Phys. Chem. B*, 111:2675–2687, 2007.

[32] A. Li and V. Daggett. Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. USA.*, 91:10430–10434, 1994.

[33] L. Li and E. I. Shakhnovich. Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA.*, 98:13014–13018, 2001.

[34] J. Gsponer and A. Caflisch. Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. USA.*, 99:6719–6724, 2002.

[35] N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121:415–425, 2004.

[36] I.A. Hubner, J. Shimada, and E.I. Shakhnovich. Commitment and nucleation in the protein G transition state. *J. Mol. Biol.*, 336:745–761, 2004.

[37] S.S. Cho, Y. Levy, and P.G. Wolynes. P versus Q: Structural reaction coordinates capture protein folding and smooth landscapes. *Proc. Natl. Acad. Sci. USA.*, 103:586–591, 2006.

[38] G. Settanni and A. Fersht. High temperature unfolding simulations of the TRPZ1 peptide. *Biophys. J.*, 94:4444–4453, 2008.

Chapter 2

Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein.

[*Proteins: Structure, Function and Bioinformatics*, **2008**, *70(4)*,1185-1195]

WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein

Stefanie Muff and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

ABSTRACT

The effects of a single-point mutation on folding thermodynamics and kinetics are usually interpreted by focusing on the native structure and the transition state. Here, the entire conformational spaces of a 20-residue three-stranded antiparallel β-sheet peptide (double hairpin) and of its single-point mutant W10V are sampled close to the melting temperature by equilibrium folding–unfolding molecular dynamics simulations for a total of 40 μs. The folded state as well as the most populated free energy basins in the denatured state are isolated by grouping conformations according to fast relaxation at equilibrium. Such kinetic analysis provides more detailed and useful information than a simple projection of the free energy. The W10V mutant has the same native structure as the wild type peptide, and similar folding rate and stability. In the denatured state, the N-terminal hairpin is about 20% more structured in W10V than the wild type mainly because of van der Waals interactions. Notably, the W10V mutation influences also the van der Waals energy at the transition state ensemble causing a shift in the ratio of fluxes between two different transition state regions on parallel folding pathways corresponding to nucleation at either of the two β-hairpins. Previous experimental studies have focused on the effects of denaturant-dependent or temperature-dependent changes in the structure of the denatured state. The atomistic simulations show that a single-point mutation in the central strand of a β-sheet peptide results in remarkable changes in the topography of the denatured state ensemble. These changes modulate the relative accessibility of parallel folding pathways because of kinetic partitioning of the denatured state. Therefore, the observed dependence of the folding process on the starting ensemble raises questions on the biological significance of in vitro folding studies under strongly denaturing conditions.

Proteins 2008; 70:1185–1195.

Key words: complex network; non-native interactions; transition state; multiple folding pathways; free-energy surface.

INTRODUCTION

Despite significant advances in the understanding of the protein folding process,[1–8] the link between primary structure and folding thermodynamics and kinetics is not completely clear. In particular, the full elucidation of the mechanism(s) of protein folding requires an in-depth understanding of the denatured state that is the starting ensemble. At physiological (i.e., mainly folding) conditions the denatured state is not only elusive to experimental characterization but also complex and highly heterogeneous even for small proteins and structured peptides.[9–12] Theoretical models and computational approaches have been developed to try to capture the complexity of the free energy surface which governs protein folding.[1,10–13] Evidence from experimental[14] and computational studies[15] indicates that the unfolded population is not featureless and can retain native-like topology, even at high concentration of denaturant according to residual dipolar couplings nuclear magnetic resonance (NMR) measurements.[16] Recently, several biophysical studies have focused on misfolding and the denatured state because of their role in protein-aggregation diseases.[17] As an example, conformers with non-native aromatic clusters have been suggested to play a role in the initiation of amyloidosis from the acid-unfolded state of β_2-microglobulin according to NMR experiments coupled with site-directed mutagenesis.[18]

The main motivation of the present simulation study was to investigate the influence of a single-point mutation on the conformational space and folding pathways of a structured peptide. It was decided to investigate a designed three-stranded antiparallel β-sheet peptide of 20 residues[19] because

Abbreviations: CSN, conformation space netwrok; TSE, transition state ensemble

This Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.

Grant sponsor: Swiss National Science Foundation.

*Correspondence to: Amedeo Caflisch, Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.
E-mail: caflisch@bioc.unizh.ch

Received 6 February 2007; Revised 10 April 2007; Accepted 16 April 2007
Published online 10 September 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21565

it folds reversibly to the correct NMR structure in molecular dynamics simulations[20–22] with a very efficient implicit solvent model.[23] The sequence of the wild type peptide is Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17}-Ile_{18}-Tyr_{19}-Thr_{20} where the Gly-Ser segment at positions 6–7 and 14–15 promote the formation of tight turns resulting in a three-stranded β-sheet consisting of two β-hairpins "sharing" the central strand.[19] The investigated mutant is Trp10Val (abbreviated, using the single-letter code of the amino acids, as W10V hereafter). In long molecular dynamics runs with several folding and unfolding events, the complete information about the native and non-native free energy basins is present in the trajectory. Therefore, peptide conformations can be grouped into free energy minima according to the equilibrium dynamics (kinetic grouping analysis). In the present study, the conformations are sampled close to the melting temperature by simulations totaling 20 μs for each peptide with more than 100 folding and unfolding events. The kinetic analysis reveals a switch of folding pathways, that is, folding begins mainly from the C-terminal and N-terminal hairpin for the wild type and the single-point mutant W10V, respectively, which is consistent with the differences in the denatured state ensemble of the two peptides. This consistency is nontrivial and originates from the very slow equilibration in the denatured state of Beta3s and W10V. In fact, analysis of the molecular dynamics trajectories provides evidence that barriers separating the free-energy basins in the denatured state ensemble are higher than those between individual non-native basins and the folded state, which is a clear indication of kinetic partitioning of the denatured state.

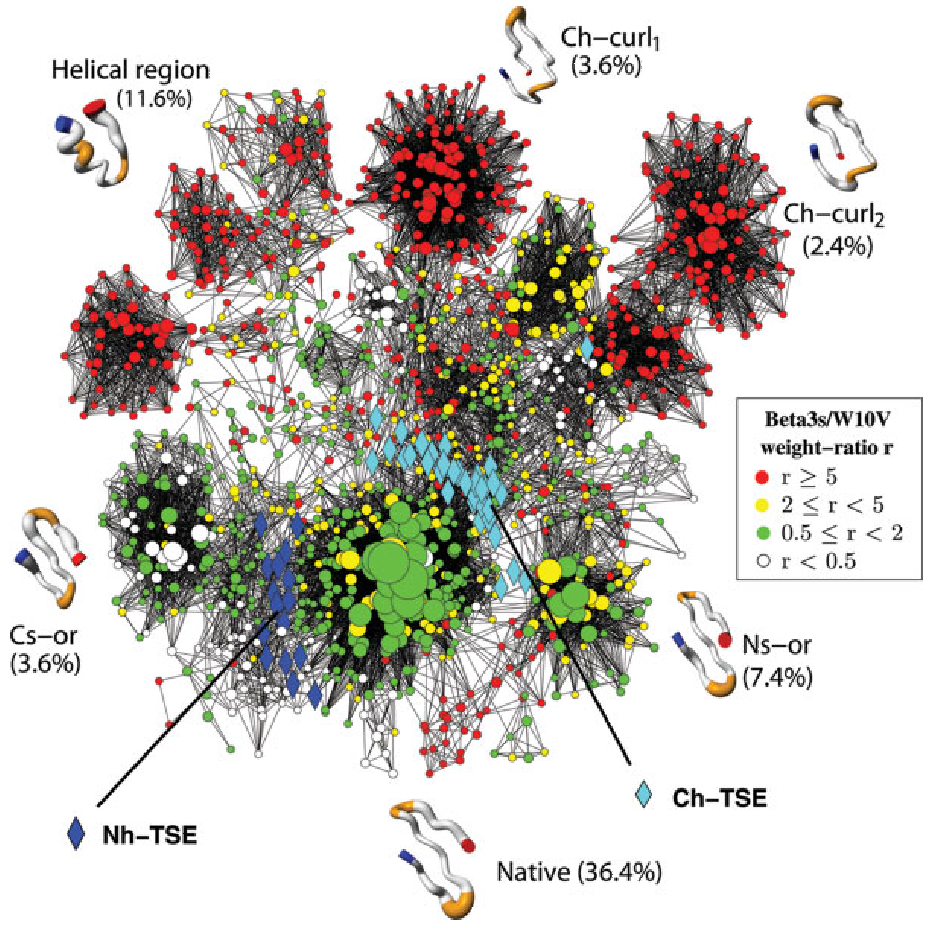


Figure 1

The CSN of Beta3s. Each node (i.e., conformation) of the network represents a secondary structure string. The surface of each node is proportional to its statistical weight and only the 1430 nodes with at least 40 snapshots for Beta3s are shown to avoid overcrowding. Nodes are colored according to the Beta3s/W10V mutant weight-ratio. Conformations in the most populated basins are shown by flexible tubes of variable diameter with N-terminus in blue, C-terminus in red and residues Gly6, Ser7, Gly14, and Ser15, which are at the two turns in the folded structure, in orange. Names of conformations are explained in the legend of Table I. Blue and cyan diamonds emphasize TSE nodes with N-terminal and C-terminal hairpin formed, respectively. This Figure was made using visone (www.visone.de) and MOLMOL.[27]

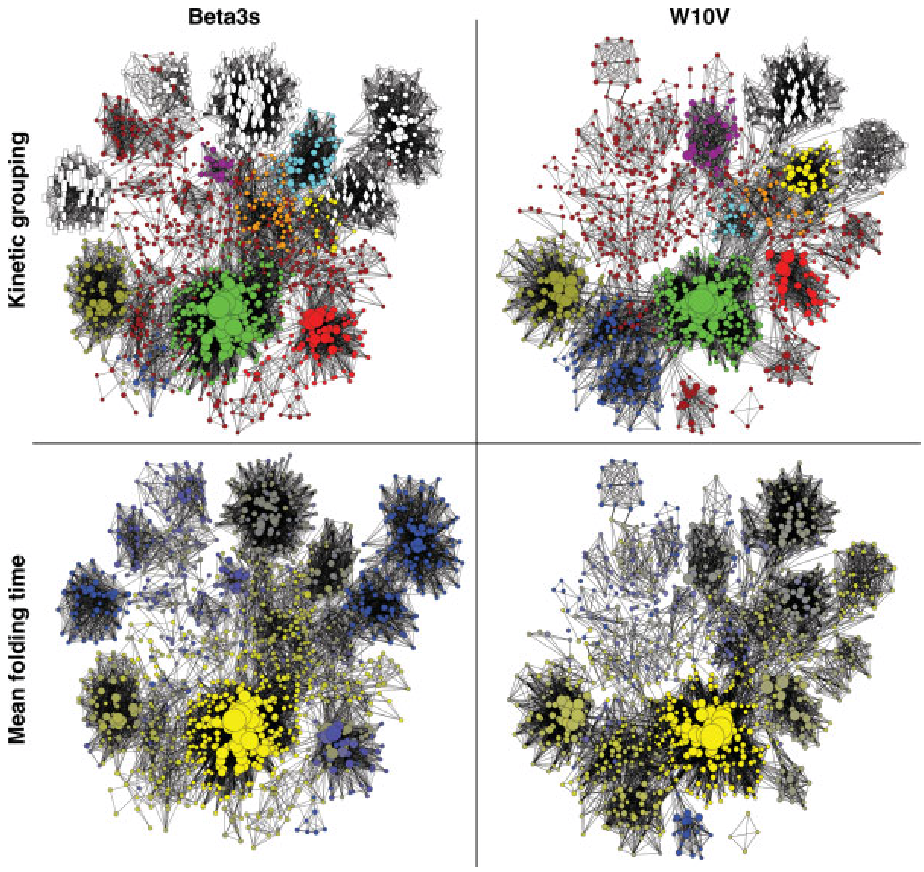


Figure 2

The CSN of Beta3s (left) and W10V (right). (Top) The coloring scheme of the kinetic grouping is chosen such that a large basin with the same most populated node in Beta3s and W10V has the same color in both CSNs (last column in Table 1). White nodes are used for basins populated significantly in only one of the two peptides. Nodes belonging to less populated basins are in brown. (Bottom) The coloring according to the mean folding time (τ_f), which is the kinetic distance from the native state, changes continuously from yellow (folding time of 0 ns) through olive (about 100 ns) to blue (200 ns). Basins identified by the kinetic grouping appear as regions of homogeneous color.

Parallel folding pathways and switches have been already observed[24,25] using protein-engineering Φ-value analysis,[26] but those studies did not report on changes in the denatured state ensemble, which is taken as the reference state in Φ-value analysis. Denaturant-dependent and temperature-dependent changes in the structured denatured state of small helical proteins have been reported.[14] The present simulation results go beyond the available experimental observations by unmasking the consequences of a hydrophobic side chain mutation on both the denatured state ensemble and transition state conformations of a β-sheet peptide.

RESULTS

The following analysis is based on a total simulation time of 20 μs at 330 K for each peptide. The wild type peptide (called Beta3s henceforth) visits 262,433 conformations, that is, nodes of the conformation space network (CSN,[11] see Methods), with a total of 534,383 direct transitions (i.e., CSN links) among nodes. The CSN of the W10V mutant is made up of 245,032 nodes and 476,721 links. A total of 120 and 105 folding events, that is, visits to the native node, were observed for Beta3s and W10V, respectively, with an average folding time of about 0.1 μs for both peptides.

Native state basin

The most populated conformation of Beta3s and its W10V mutant has a statistical weight of 5.6 and 8.8%, respectively. It is the three-stranded antiparallel β-sheet with type II′ turns at residues 6–7 and 14–15 (secondary structure string -EEEESSEEEEEESSEEEE-) in agreement with NMR data.[19] Using this conformation as "attractor," the kinetic grouping (see Methods and Suppl. Mat. for the

illustrative application of kinetic grouping to the alanine dipeptide) yields a native-basin weight of 36 and 41% for Beta3s and W10V, respectively. The CSN of Beta3s is a useful illustration of the underlying free energy surface and its dynamic connectivity. It is colored according to the Beta3s/W10V ratio of the statistical weight of individual nodes (i.e., conformations) in Figure 1, while the corresponding figure for W10V is included in the Suppl. Mat (Fig. S1). Individual nodes in the native state basin have very similar statistical weight for both peptides. As an example, there is only a 1% probability (weighted by the number of snapshots in a node and considering nodes visited at least 10 times during the total simulation time of 40 μs) that the Beta3s/W10V weight-ratio of a native-basin node is larger than 5 or smaller than 1/5, which corresponds to an absolute value of the free energy difference larger than 1 kcal/mol at 330 K. Furthermore, an average value of only 2.1 ± 1.4 Å is obtained for the root mean square deviation of C_α carbon coordinates measured on the 10^8 pairs of structures from subsets of 10^4 native state snapshots for each peptide. The corresponding values within the native basin of Beta3s and W10V are 2.6 ± 1.7 Å and 1.4 ± 0.6 Å, respectively. Hence, the folded state of the two peptides is essentially identical.

Denatured state ensemble

The two peptides show significant differences in the denatured state ensemble (Figs. 1 and 2 top). There is a 33% probability (weighted as explained above) that the Beta3s/W10V weight-ratio of a denatured-state node is larger than 5 (red nodes in Fig. 1) or smaller than 1/5. This result is not affected by statistical error (Suppl. Mat. Section I.). Additional evidence of the large discrepancies in the denatured state ensemble is provided by the CSNs of Beta3s and W10V colored according to the basins identified by the kinetic analysis (Fig. 2 top).

The mean folding time (τ_f) is the kinetic distance from the native state, defined as the time interval Δt between a snapshot and the most populated node, where Δt is averaged over all the snapshots in a given node. By coloring the nodes according to the value of τ_f the basins identified by the kinetic grouping emerge as network "regions" of homogeneous colors (compare top and bottom of Fig. 2). This observation validates the kinetic grouping analysis if one considers that relaxation times to the native node were not used for grouping.

Free energy basins

The changes in the denatured state ensemble due to the W10V mutation are one of the main results of the kinetic grouping (Table I). There is a striking difference between the detailed information provided by the kinetic grouping (and illustrated by the CSN) and the free energy profile along the number of native contacts (thick lines in Fig. 3). The latter is a projection which yields a very similar picture of the denatured state of Beta3s and W10V even if the native state is used as reference for both profiles (as in Fig. 3). The complexity of the denatured state of Beta3s and W10V is captured by kinetic grouping analysis. The enthalpic basins identified by ki-

Table I
Results of the Kinetic Grouping

		Beta3s				W10V				
Conformation	Name	Weight (%) node	basin	τ_f (ns) node	basin	Weight (%) node	basin	τ_f (ns) node	basin	Color
-EEEESSEEEEEESSEEEE-	native	5.6	36.4	—	—	8.8	40.6	—	—	Green
Larger weight in Beta3s										
-EEEESTTEEEEESSEEEE-	Ns-or	1.2	7.4	138	109	0.8	4.0	92	90	Red
---SSGGG---EESSEETT-	Ch-curl$_1$	0.1	3.6	98	90	0	0	—	—	White ovals
---SSGGG-EESSTTTTEE-	Ch-curl$_2$	0.1	2.4	285	257	0	0	—	—	White circles
-----SS--EEEESSEEEE-		0.03	2.2	53	75	0.04	0.9	72	85	Orange
-HHHHHHHHHHHHS------	Helix$_{1-13}$	0.1	2.1	137	122	0.01	0.5	124	151	White squares
--EESSSEEEEEESSEEEE-		0.1	1.9	87	84	0.04	0.6	148	134	Cyan
---SSGGG-EEESSSEEEE-		0.1	1.2	200	198	0	0	—	—	White rectangles
---SSSS--EEESTT-EEE-		0.06	0.9	316	263	0	0	—	—	White diamonds
Larger weight in W10V										
-EEEESSEEEEESSSEEEE-	Cs-or	0.4	3.6	63	70	0.7	6.0	69	75	Olive
-EEEESSEEEE---TT--B-	Nh-curl	0.04	0.5	59	58	0.1	3.2	69	69	Blue
----STT---EEESSEEEE-		0.1	0.8	139	113	0.3	2.7	108	121	Violet
--BSS-SSSEEE-STTEEE-	Ch-curl$_3$	0	0	—	—	0.1	2.6	104	105	White diamonds
--SSSSS--EEEESSEEEE-		0.03	0.8	103	97	0.1	2.1	111	100	Yellow
-BSSSS---EEEESSEEEE-		0	0	—	—	0.02	0.3	61	53	White circles

Statistical weight of the native basin and the most populated free energy basins in the denatured state as identified by the kinetic grouping. Mean folding times (τ_f) are average values for snapshots in a node or basin. Conformations with names are shown by flexible tubes of variable diameter in Figure 1 for Beta3s and in the Suppl. Mat. (Fig. S1) for W10V: Ns-or, N-terminal strand out of register; Cs-or, C-terminal strand out of register; Nh-curl, curl-like conformation with structured N-terminal hairpin; Ch-curl, curl-like conformation with structured C-terminal hairpin. The colors indicated in the last column are those used in Figure 2 top.

DOI 10.1002/prot

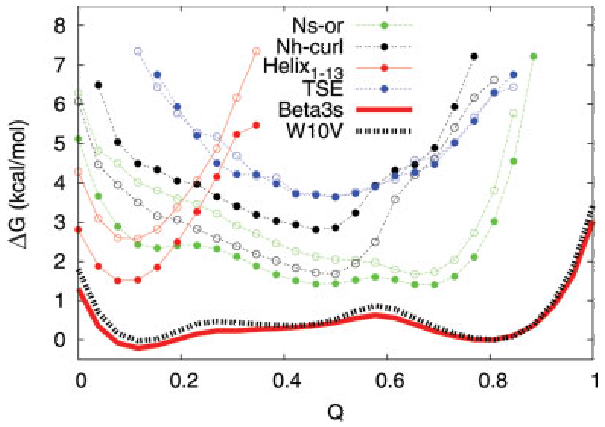


Figure 3

Inadequacy of free energy projection. The free energy profile along the fraction of native contacts Q (see[20] for the list of contacts) is plotted with thick lines without symbols. The contributions of individual free energy basins in the denatured state are shown with thin lines with solid and empty circles for Beta3s and W10V, respectively. The folded state basin is not shown to avoid overcrowding; it overlaps with the thick lines in the range 0.7 ≤ Q ≤ 1.0. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

netic grouping and the entropically favored helical region (see below and Ref. 11) can be plotted as a supplement to the free energy profile (thin lines with symbols in Fig. 3) to demonstrate the inadequacy of simple one-dimensional projections.

Notably, the basin with helical N-terminal segment and the β-sheet with N-terminal strand out of register (Ns-or) is more populated in Beta3s than W10V, whereas the opposite is observed for the curl-like basin with stable N-terminal hairpin (Nh-curl, see Table I for names of most populated free energy basins). Moreover, the kinetic grouping analysis shows that two of the three most populated free energy basins in the denatured state of W10V have a native-like N-terminal hairpin; they are the β-sheet with C-terminal strand out of register (Cs-or) and the Nh-curl conformation (Table I). Overall, the segment 2–11 has 19% higher content of native secondary structure and 20% more native tertiary contacts (Suppl. Mat. Fig. S5) in the denatured state of the W10V mutant than the wild type peptide.

Analysis of the interaction energy between residue pairs in Cs-or and Nh-curl conformations shows that the loss of favorable interactions of residue 10 with Lys17 and Tyr19 destabilizes Beta3s more than W10V mainly because of the difference in van der Waals interactions (Fig. 4). This loss is compensated, but only partially, in the Cs-or conformations, where the shift of the β-strand register results in non-native van der Waals interactions of residue 10 with Thr16 and Ile18, which are more favorable for the Trp10 side chain of the wild type Beta3s than the smaller Val10 of the mutant. These mainly hydrophobic contacts involving Trp10 provide further evidence for the importance of non-native interactions and are consistent with experimental observations. In fact, residual structure in the denatured state of an SH3 domain and non-native interactions of its Trp36 side chain have been observed by NMR spectroscopy under folding conditions.[28] Interestingly, the native structure of the SH3 domain consists mainly of a hairpin packed against a three-stranded antiparallel β-sheet, and Trp36 is in the central strand of the latter.

Conformations with partial helical content

It was shown in a previous study[11] that the denatured state ensemble of Beta3s is very heterogeneous and

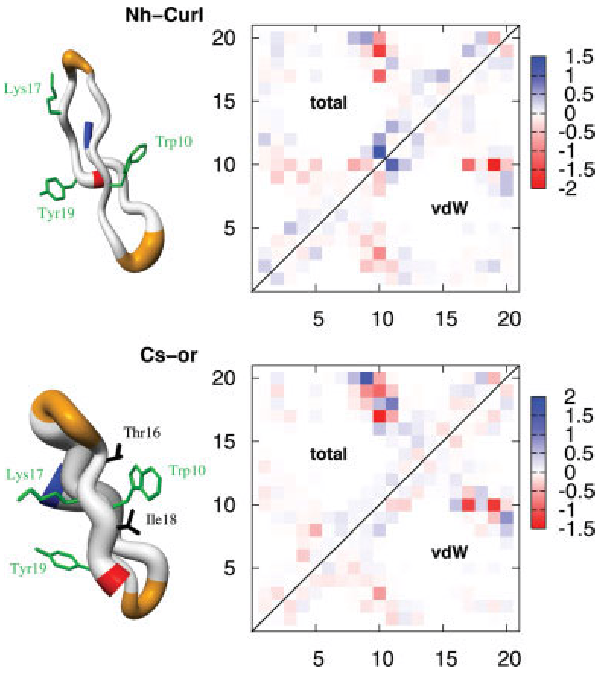


Figure 4

The non-native conformations Nh-curl (top) and Cs-or (bottom) are more populated in W10V than Beta3s. (Left) Flexible tube representation: side chains which interact favorably with residue 10 in the native state but are distant from it in the Nh-curl and Cs-or conformations (see text) are in green. Side chains involved in non-native contacts with Trp10 are in black. (Right) Difference in residue pairwise interaction energy between W10V and Beta3s averaged over the basin. The pairwise energy values in the native state are used as reference and subtracted from all values in the matrix. A red square indicates that the corresponding pair of residues has a more favorable energy in the Nh-curl or Cs-or conformation of W10V than Beta3s. The bars on the right give values in kcal/mol. The upper and lower triangular matrices show the total (sum of van der Waals and electrostatics) and van der Waals energy, respectively, and their similarity indicates that most of the enthalpic stabilization originates from differences in van der Waals energy. In fact, the sum of all pairwise contributions yields a W10V vs. Beta3s difference of −2.0 kcal/mol for the Nh-curl conformation, with a van der Waals contribution of −3.9 kcal/mol. The corresponding values for the Cs-or conformation are −0.2 kcal/mol and −2.2 kcal/mol.

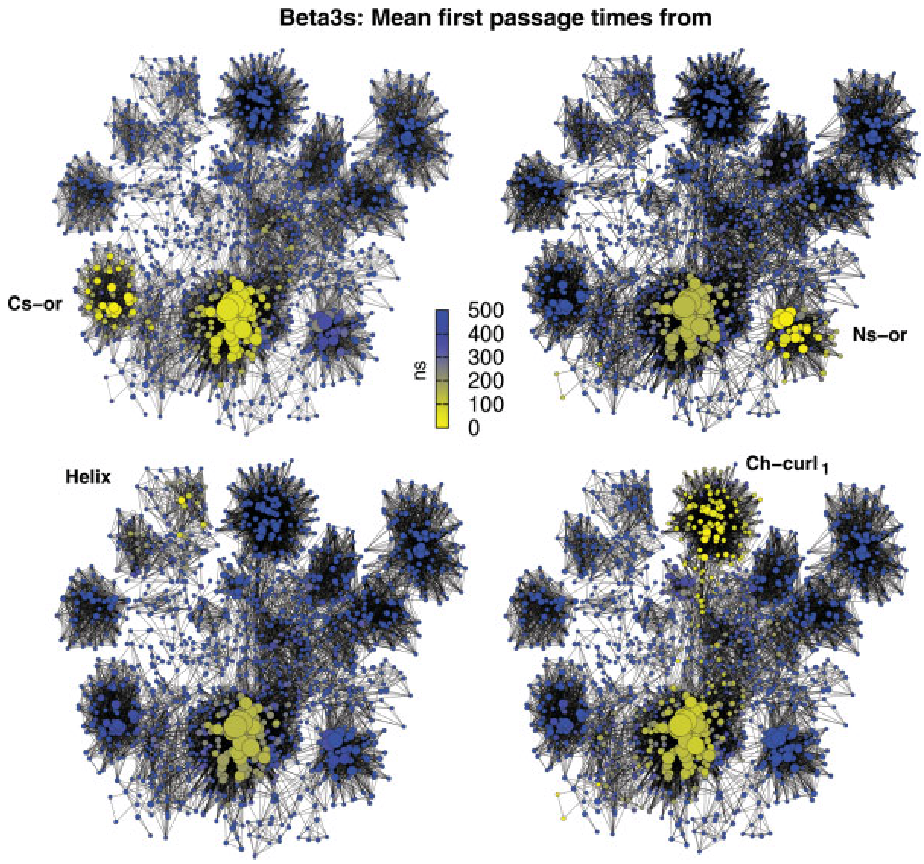


Figure 5

The denatured state is kinetically partitioned. Nodes are colored according to mean first passage times from the most populated node of individual free energy basins to all other nodes of the CSN of Beta3s. Nodes within the basin of the starting node (i.e., the most populated node of the considered basin) are visited relatively fast (yellow), indicating rapid intrabasin transitions and supporting the kinetic grouping analysis (Fig. 2 top). Equilibration between different unfolded basins (blue) is slower than reaching the folded state (olive) which shows that the denatured state is kinetically partitioned, i.e., no fast equilibration takes place between basins in the denatured state. In other words, the native state is a hub.[11] Kinetic partitioning is also observed for the denatured state of W10V (Suppl. Mat. Fig. S3). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

includes conformations with partial helical structure and fluctuating unstructured residues, for example, the C-terminal segment in the helical conformation shown on top left of Figure 1. Here, it is interesting to compare the helical propensity of the two peptides. The conformations of Beta3s and W10V with partial helical content (defined as more than three, four, or five consecutive 3_{10}-, α-, or π-helical residues, respectively) have a statistical weight of 11.6 and 10.7%, respectively. Note that the aforementioned basin with helical N-terminal segment ($Helix_{1–13}$, -HHHHHHHHHHHHS------) is only a fraction (of weight 2.1 and 0.5% for Beta3s and W10V, respectively, Table I) of the conformations with partial helical structure. The latter make up an entropic region in the denatured state.[11] The weight distribution of the nodes in the helical region shows a more pronounced decay than for the native state nodes (Suppl. Mat. Fig. S12), which is consistent with the entropic character of the helical region. In fact, it was demonstrated recently on low-dimensional models, which can be treated analytically, that free energy basins with entropic character have a faster decay of the weight distribution than mainly enthalpic basins.[29]

Interestingly, the percentage of helical content along the sequence shows that Beta3s is more helical than W10V in the central segment, that is, residues 7–13 (Suppl. Mat. Fig. S11). This observation is consistent with the experimental finding that the side chain of valine has a destabilizing effect on the helical structure[30] because of the branching at the C_β carbon.

Kinetic partitioning

The denatured state ensemble of Beta3s and W10V is not in rapid equilibrium. In fact, the mean first passage times (mfpt) from the most populated node of individual

basins in the denatured state to all other nodes, which can be calculated by simply following the trajectories (i.e., the time series of the nodes), reveal that the fastest transitions are those to the native basin. It is useful to color the CSN of Beta3s according to values of mpft from individual free-energy basins (Fig. 5). Folding from four of these basins take in average less than 150 ns (olive nodes in Fig. 5), while the mean transition times to other regions are larger than 500 ns (blue nodes in Fig. 5), indicating that trajectories between non-native basins mostly pass through the folded state. A similar picture is observed for W10V (Suppl. Mat. Fig. S3). The barriers between individual free energy basins in the denatured state and the native state are clearly lower than barriers within the unfolded state. The observation of kinetic partitioning of the denatured state by kinetic grouping analysis provides further evidence to the fact that the native basin of Beta3s and W10V acts as a hub.[11] Importantly, these simulation results are consistent with recent experimental reports on competing folding routes and kinetic partitioning in ubiquitin[31] and the DNA-binding domain of p53.[32] Most notably, the present results provide evidence that folding pathways depend crucially on the denatured state, that is, the starting ensemble. Therefore, *in vitro* folding experiments under strongly unfolding conditions (e.g., high concentration of chemical denaturants) do not necessarily give insights into the folding process under physiological conditions.

Transition state ensemble and switch of folding pathways

The folding probability P_{fold} of a conformation is the likelihood to reach the most populated node before unfolding.[33] It should be close to 0.5 for transition state ensemble (TSE) structures. A total of 55 nodes containing 1703 snapshots for Beta3s and 51 nodes with 1881 snapshots for W10V (diamonds in Fig. 1 and Fig. S1) have been isolated as TSE by the node-P_{fold} analysis.[34] The native secondary structure content in a node can be determined by comparison to the native string. Figure 6 shows the percentage of native secondary structure along the sequence for the TSE of Beta3s and W10V. Notably, there is a TSE switch from predominance of structured C-terminal hairpin in Beta3s to structured N-terminal hairpin in W10V, which is consistent with the redistribution of basin populations in the denatured state ensembles because of the kinetic partitioning, that is, higher barriers between non-native free-energy basins than folding barriers. Furthermore, each TSE string can be unambiguously classified into structured N-terminal (Nh-TSE) or C-terminal (Ch-TSE) hairpin by predominance of native bits in the 2–11 or 10–19 string segment, respectively. In the TSE of Beta3s, 70% of the structures belong to Ch-TSE and only 30% to Nh-TSE, whereas W10V

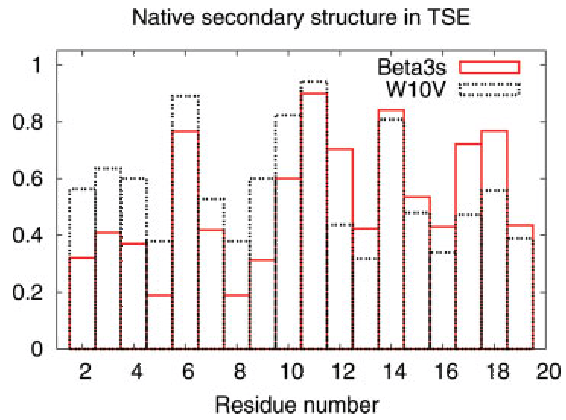


Figure 6

The content of native secondary structure, measured by DSSP,[35] at the TSE illustrates the shift of flux. The pathway through the TSE with formed N-terminal hairpin is more populated in W10V, whereas the pathway through the C-terminal hairpin is more populated in Beta3s. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

switches to 45% Ch-TSE and 55% Nh-TSE. These results are robust with respect to the choice of a τ_{commit} value in the interval from 1.5 to 5 ns (see Suppl. Mat. Fig. S8).

Interestingly, an energetic analysis explains the switch of flux through the two distinct TSEs. The effective energy contribution (sum of intrapeptide and solvation energy terms) to the Nh-TSE activation energy is 0.6 kcal/mol less favorable for W10V than Beta3s. This value is smaller than the effective energy contribution to the Ch-TSE which is 1.6 kcal/mol less favorable for W10V than Beta3s. Therefore, if one assumes similar conformational entropy contributions for both peptides, it follows that the W10V mutation raises more the Ch-TSE than Nh-TSE barrier with respect to the wild type barriers. Most of these energetic effects at the TSE are due to van der Waals interactions as shown in the Suppl. Mat. (Fig. S9).

The free energy projection along the fraction of native contacts shows again its limitations. In fact, using a "naive" criterion of about half of the native contacts formed it is much more likely to select Ns-or conformations than TSE structures (Fig. 3). The former belong to the denatured state ensemble and are clearly pre-critical as indicated by their node-P_{fold} values very close to zero (between 0.001 and 0.005).

CONCLUSIONS

A kinetic analysis based on coarse-graining of molecular dynamics snapshots and grouping according to fast relaxation at equilibrium has been presented for investigating the topography of the conformational space of structured peptides at the melting temperature. Applica-

tion to simulations of reversible folding of a three-stranded antiparallel β-sheet peptide of 20 residues reveals that the W10V mutation in the central strand alters the relative population of the non-native states mainly because of differences in van der Waals energy. In particular, changes are observed in the denatured state ensemble, which is found to be kinetically partitioned, as well as in the relative height of two distinct transition state barriers on parallel folding pathways. As a consequence, the minor folding route in the wild type (first formation of N-terminal hairpin) becomes the major route in the W10V mutant. In other words, the point mutation modulates the flux through each pathway. It is generally accepted that the final stage of folding requires specific side chain packing. The present study provides evidence that a single side chain difference can result in a redistribution of basin populations in a structured denatured state ensemble, that is, at the beginning of the folding process. Significant differences have been observed in the denatured state of a small protein under mildly and strongly denaturing conditions by a joint analysis of fluorescence, differential scanning calorimetry, circular dichroism, and NMR spectroscopy.[36] Therefore, the folding mechanism and pathways postulated on the basis of *in vitro* experiments with high concentrations of guanidinium hydrochloride or urea might not be representative of the folding process under physiological conditions. Further experimental and computational investigations on denatured states under physiological conditions are required because the "starting ensemble" for folding plays a key role and is not just a randomly unfolded polypeptide chain.

Recent network and graph analyses have shown that the surprisingly simple two-state picture of protein folding, often obtained by projecting the free energy onto an arbitrarily chosen progress variable, is not consistent with the complexity of the actual free energy surface.[10–12,37] The simulation results and kinetic analysis provide not only a detailed description of the heterogeneity of the denatured state ensemble, but capture also the changes originating from a single-point mutation, which are hidden in simple projections of the free energy. The kinetic grouping approach presented here is quite general and can be used for investigating the free energy topography of other complex (molecular) systems provided that their conformation space can be sampled by equilibrium simulations.

METHODS

Molecular dynamics simulations

All simulations and part of the analysis of the trajectories were performed with the program CHARMM.[38] The designed 20-residue peptide Beta3s[19] and its W10V mutant were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[38] with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent on the solute.[23] This choice is justified by two reasons. First, using explicit water simulations it is not possible to sample a statistically significant number of equilibrium folding–unfolding transitions which is a *conditio sine qua non* for the present analysis. Second, it was shown previously using exactly the same SAS-based implicit solvation model that at 330 K Beta3s folds reversibly to its NMR conformation irrespective of the starting structure, and importantly, 23 of the 26 NOE restraints are satisfied.[20] Despite the absence of collisions with water molecules, in the simulations with implicit solvent relative rates are comparable with the values observed experimentally. Helices fold in about 1 ns,[39] β-hairpins in about 10 ns,[39] and triple-stranded β-sheets in about 0.1 μs,[22] while the experimental values are ~0.1 μs,[40] ~1 μs[40] and ~10 μs,[19] respectively. Importantly, the small variations in total solvent accessible surface and radius of gyration during folding of Beta3s at 330 K[21] provide evidence that the lack of solute/solvent friction does not have a significant effect on the main results of the kinetic analysis.

For each peptide 10 molecular dynamics runs of 2 μs each with different initial distribution of velocities were performed with the Berendsen thermostat (coupling constant of 5 ps) at 330K, which is close to the melting temperature of wild-type Beta3s.[21] A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 10^6 snapshots for each system. The significant amount of sampling is supported by the small variation in the native population measured on two disjoint 10-μs subsets of the trajectories that have a weight of 38 and 35% for Beta3s (38 and 43% for two subsets of the W10V trajectories). Moreover, a small variation is observed for the population of individual free energy basins in the denatured state. As an example, the Cs-or basin has a weight of 3.9 and 3.2% for the two 10-μs subsets of Beta3s (6.2 and 5.9% for W10V), and the population of the Nh-curl basin is 0.35 and 0.65% for the two 10-μs subsets of Beta3s (3.4 and 3.1% for W10V).

Coarse-graining

There are several ways for assigning snapshots (i.e., coordinate sets) to coarse-grained conformations (nodes and strings are used as synonyms in this article) and different types of analysis might require different coarse-graining approaches.[11,12] The coarse-graining used in this work is based on secondary structure strings.[35] A conformation is a single string of secondary structure, for example, the most populated conformation of Beta3s is -EEEESSEEEEEESSEEEE-.[11] There are 8 possible

 DOI 10.1002/prot

"letters" in the secondary structure "alphabet": "H", "G", "I", "E", "B", "T", "S", and "−", standing for α helix, 3_{10} helix, π helix, extended, isolated β-bridge, hydrogen bonded turn, bend, and unstructured, respectively. Since the N- and C-terminal residues are always assigned an "−",[35] a 20-residue peptide can in principle assume $8^{18} \simeq 10^{16}$ conformations. The secondary structure-based coarse-graining is appropriate for structured peptides without loops and has three advantages with respect to approaches based on root mean square deviation (RMSD) of atomic coordinates. First, it does not require the use of an arbitrarily chosen threshold value. Second, each node is uniquely defined by its secondary structure string which is a useful conformational "label". Third, the same type of secondary structure classification[35] is used for coarse-graining and analysis of structural properties (e.g., native content in TSE, Fig. 6).

Conformation space network

Conformations (i.e., secondary structure strings) are nodes of the conformation space network (CSN) and the direct transitions between them are links.[11] The number of snapshots with a given secondary structure string is abbreviated as $\tilde{w}$. The statistical weight w of a node is equal to $w = \tilde{w}/N$, where $N = 10^6$ is the total number of snapshots for each of the two peptides. The CSN in Figure 1 and in the Suppl. Mat. (Fig. S1) show only nodes with $\tilde{w} \geq 40$ to avoid overcrowding. It is important to emphasize that the CSN of "heavy" nodes is used solely for illustrative purposes, whereas the quantitative analysis was performed using the kinetic grouping which takes into account the complete trajectory, that is, all nodes (see later).

Kinetic grouping into free energy basins

An important methodological aspect of the present work is that conformations are grouped into disjoint free energy basins not according to arbitrarily chosen geometric characteristics, but rather according to the statistics of the transitions at equilibrium as explained in the two following subsections.

The native state basin

The log-binned distribution of first passage times from any snapshot saved along the trajectory to the native node -EEEESSEEEEEESSEEEE- separates the conformational space into two regions: the fast relaxation within the native basin takes place in less than 1.6 ns, while folding from outside is about two orders of magnitudes slower (Fig. 7). Nodes are assigned to the native basin if they have a node-$P_{fold} \geq 0.5$ (see Ref. 34) using a commitment time τ_{commit} of 1.6 ns, meaning that at least 50% of the structures in a node must visit the native node within 1.6 ns along the trajectory. Note that this is a purely kinetic argument for grouping nodes to a basin. The only condition is that the underlying coarse-graining yields nodes containing kinetically homogeneous structures.

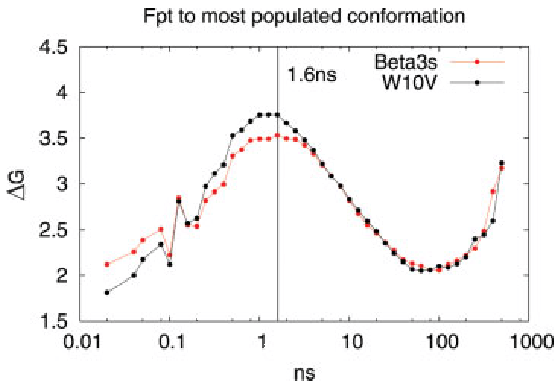


Figure 7

Distribution of first passage times P(fpt) to the most populated node. Values are calculated as $\Delta G = -k_B T \ln(P(fpt))$ *and given in kcal/mol. The* fpt *can be considered as a geometrically unbiased reaction coordinate. This projection is very useful to determine the transition between intra-basin and inter-basin relaxation, which is emphasized by the logarithmic binning (10 bins/decade) without normalization of the binsize. The peak at about 0.1 ns is an artifact of the binning but does not have any effect on the results. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

Since the accuracy of node-P_{fold} depends on the population of the node, a statistical argument to reduce the number of false negatives is used. If a node has $P_{fold} \geq 0.5$ in the limit of infinite sampling, it should be detected with a probability of at least 80% (called "*80% criterion*" hereafter). As an example, a node with 20 snapshots is grouped to the native basin if at least 8 snapshots are separated from the native node by less than 1.6 ns (node-$P_{fold} \geq 0.4$), while a node with 200 snapshots must have at least 92 of them folding within 1.6 ns (node-$P_{fold} \geq 0.46$).

Basins in the denatured state

The kinetic grouping can be extended to automatically detect the attractor regions which are stabilized mainly enthalpically. Instead of calculating node-P_{fold}, which per definition follows the trajectory to the native state, an analogous quantity can be determined between any two nodes. Hence, $P_{commit}(i \rightarrow j)$ is defined as the probability to observe a transition from node i to node j within a given commitment time τ_{commit}. The value of $P_{commit}(i \rightarrow j)$ is an asymmetric, directed measure of the kinetic similarity of nodes i and j. Once the P_{commit}-matrix has been calculated for nodes with at least 40 snapshots, pairs of nodes (i,j) are grouped together if $P_{commit}(i \rightarrow j) \geq 0.5$, where the statistical 80%-criterion is applied. This way of grouping leads to a partitioning into disjoint sets, that is, a disconnected CSN, whose subgraphs correspond to different attractor regions. A value of $\tau_{commit} = 1$ ns is used

because it is in the range of the relaxation times for most of the important basins, but still two to three orders of magnitudes shorter than typical transition times from outside (see Suppl. Mat. Fig. S6). Note that the kinetic grouping is robust with respect to the choice of the commitment time in the interval $0.5 \leq \tau_{commit} \leq 2.0$ ns (Suppl. Mat. Fig. S7). Moreover, different values of τ_{commit} allow one to analyze different levels of ruggedness of the free energy surface as illustrated by the application to the alanine dipeptide (Suppl. Mat.). Since the all-against-all P_{commit} calculations are expensive (growth is quadratic) and to reduce noise caused by nodes lying in high-energy regions, only nodes with at least 40 snapshots were used in a first step (which required about 10 h of CPU time of a Xeon 2 GHz), while the small nodes were grouped to the attractor regions in a post-processing step as detailed in the Suppl. Mat.

Transition state ensemble

Putative transition state ensemble (TSE) nodes are identified by the node-P_{fold} analysis[34] using the aforementioned 80%-criterion and with a τ_{commit} of 1.6 ns to be consistent with the native state basin isolation. Only nodes with weight above a threshold ($\tilde{w} \geq 20$) are considered to guarantee a minimal amount of statistics. Validation with classical P_{fold} analysis[33] is presented in the Suppl. Mat. (Table S-III).

ACKNOWLEDGMENTS

We thank Enrico Guarnera, Francesco Rao, and Gianni Settanni for very interesting and helpful discussions. We also thank Paolo De Los Rios, David Gfeller, and Ben Schuler for a critical reading of the manuscript. We are indebted to Michele Seeber for the program Wordom[41] used to analyze the trajectories. The simulations were performed on the Matterhorn cluster of the University of Zurich and we gratefully acknowledge the support of C. Bolliger and A. Godknecht.

REFERENCES

1. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscape and motions of proteins. Science 1991;254:1598–1603.
2. Dill K, Chan H. From levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
3. Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 1992;89:8721–8725.
4. Dinner AR, Šali A, Smith LJ, Dobson CM, Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment. Trends Biochem Sci 2000;25:331–339.
5. Daggett V, Fersht AR. The present view of the mechanism of protein folding. Nat Rev Mol Cell Biol 2003;4:497–502.
6. Wolynes PG. Energy landscapes and solved protein-folding problems. Phil Trans R Soc A 2005;363:453–467.
7. Thirumalai D, Hyeon C. RNA and protein folding: common themes and variations. Biochemistry 2005;44:4957–4970.
8. Shakhnovich EI. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 2006;106:1559–1588.
9. Dill KA, Shortle D. Denatured states of proteins. Annu Rev Biochem 1991;60:795–825.
10. Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. Proc Natl Acad Sci USA 2004;101:14766–14770.
11. Rao F, Caflisch A. The protein folding network. J Mol Biol 2004;342:299–306.
12. Hubner IA, Deeds EJ, Shakhnovich EI. Understanding ensemble protein folding at atomic detail. Proc Natl Acad Sci USA 2006;103:17747–17752.
13. Boczko EM, Brooks CL III. First-principles calculation of the folding free energy of a three-helix bundle protein. Science 1995;269:393–396.
14. Ferguson N, Sharpe TD, Schartau PJ, Sato S, Allen MD, Johnson CM, Rutherford TJ, Fersht AR. Ultra-fast barrier-limited folding in the peripheral subunit-binding domain family. J Mol Biol 2005;353:427–446.
15. Zagrovic B, Snow CD, Khaliq S, Shirts MR, Pande VS. Native-like mean structure in the unfolded ensemble of small proteins. J Mol Biol 2002;323:153–164.
16. Shortle D, Ackerman MS. Persistence of native-like topology in a denatured protein in 8 M urea. Science 2001;293:487–489.
17. Dobson CM. Protein folding and misfolding. Nature 2003;426:884–890.
18. Platt G, McParland V, Kalverda AP, Homans SW, Radford SE. Dynamics in the unfolded state of β_2-microglobulin studied by NMR. J Mol Biol 2005;346:279–294.
19. De Alba E, Santoro J, Rico M, Jiménez MA. De novo design of a monomeric three-stranded antiparallel β-sheet. Protein Sci 1999;8:854–865.
20. Ferrara P, Caflisch A. Folding simulations of a three-stranded antiparallel β-sheet peptide. Proc Natl Acad Sci USA 2000;97:10780–10785.
21. Cavalli A, Ferrara P, Caflisch A. Weak temperature dependence of the free energy surface and folding pathways of structured peptides. Proteins: Struct Funct Bioinformatics 2002;47:305–314.
22. Settanni G, Rao F, Caflisch A. Φ-Value analysis by molecular dynamics simulations of reversible folding. Proc Natl Acad Sci USA 2005;102:628–633.
23. Ferrara P, Apostolakis J, Caflisch A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins: Struct Funct Bioinformatics 2002;46:24–33.
24. Nauli S, Kuhlman B, Baker D. Computer-based redesign of a protein folding pathway. Nat Struct Biol 2001;8:602–605.
25. Wright CF, Lindorff-Larsen K, Randles LG, Clarke J. Parallel protein-unfolding pathways revealed and mapped. Nat Struct Biol 2003;10:658–662.
26. Fersht AR, Matouschek A, Serrano L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. J Mol Biol 1992;224:771–782.
27. Koradi R, Billeter M, Wüthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph Modell 1996;14(1):51–55.
28. Crowhurst KA, Tollinger M, Forman-Kay JD. Cooperative interactions and a non-native buried Trp in the unfolded state of an SH3 domain. J Mol Biol 2002;322:163–178.
29. Gfeller D, De Los Rios P, Caflisch A, Rao F. Complex network analysis of free-energy landscapes. Proc Natl Acad Sci USA 2007;104:1817–1822.
30. Hecht MH, Zweifel BO, Scheraga HA. Helix-coil stability constants for the naturally occurring amino acids in water: XVII threonine parameters from poly(hydroxylbutyl-glutamine-co-L-threonine). Macromolecules 1978;11:545–551.

31. Crespo MD, Simpson ER, Searle MS. Population of on-pathway intermediates in the folding of ubiquitin. J Mol Biol 2006;360: 1053–1066.
32. Butler JS, Loh SN. Kinetic partitioning during folding of the p53 DNA binding domain. J Mol Biol 2005;350:906–918.
33. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich EI. On the transition coordinate for protein folding. J Chem Phys 1998;108:334–350.
34. Rao F, Settanni G, Guarnera E, Caflisch A. Estimation of protein folding probability from equilibrium simulations. J Chem Phys 2005;122:184901.
35. Andersen CAF, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure 2002;10:174–184.
36. Mayor U, Grossmann JG, Foster NW, Freund SMV, Fersht AR. The denatured state of engrailed homeodomain under denaturing and native conditions. J Mol Biol 2003;333:977–991.
37. Caflisch A. Network and graph analyses of folding free energy surfaces. Curr Opin Struct Biol 2006;16:71–78.
38. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983; 4:187–217.
39. Ferrara P, Apostolakis J, Caflisch A. Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. J Phys Chem B 2000;104:5000–5010.
40. Eaton WA, Munoz V, Hagen J, Jas SGS, Lapidus LJ, Henry ER, Hofrichter J. Fast kinetics and mechanisms in protein folding. Ann Rev Biophys Biomolec Struc 2000;29:327–359.
41. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A. Wordom: a program for efficient analysis of molecular dynamics simulations. Bioinformatics, in press.

Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein

SUPPLEMENTARY MATERIAL

Stefanie Muff and Amedeo Caflisch

Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Keywords: Complex network, non-native interactions, transition state, multiple folding pathways

I. THE CSN OF W10V

Fig. S1 shows the analogous of Fig. 1 (main text) for W10V. The probability that the weight-ratio of a denatured-state node visited by either peptide is larger than two, five or ten (i.e., that the node of one of the two peptides has been visited more often than two, five or ten times than in the other peptide) is 67%, 33% and 24%, respectively. These differences in weight are statistically significant because the same analysis on two subsets of Beta3s trajectories yields Beta3s(0-10 μs)/Beta3s(10-20 μs) weight-ratios larger than two, five or ten with a probability of only 39%, 18% and 9%, respectively. Analogously, the W10V(0-10 μs)/W10V(10-20 μs) weight-ratios larger than two, five or ten have a probability of only 39%, 10% and 8%, respectively.

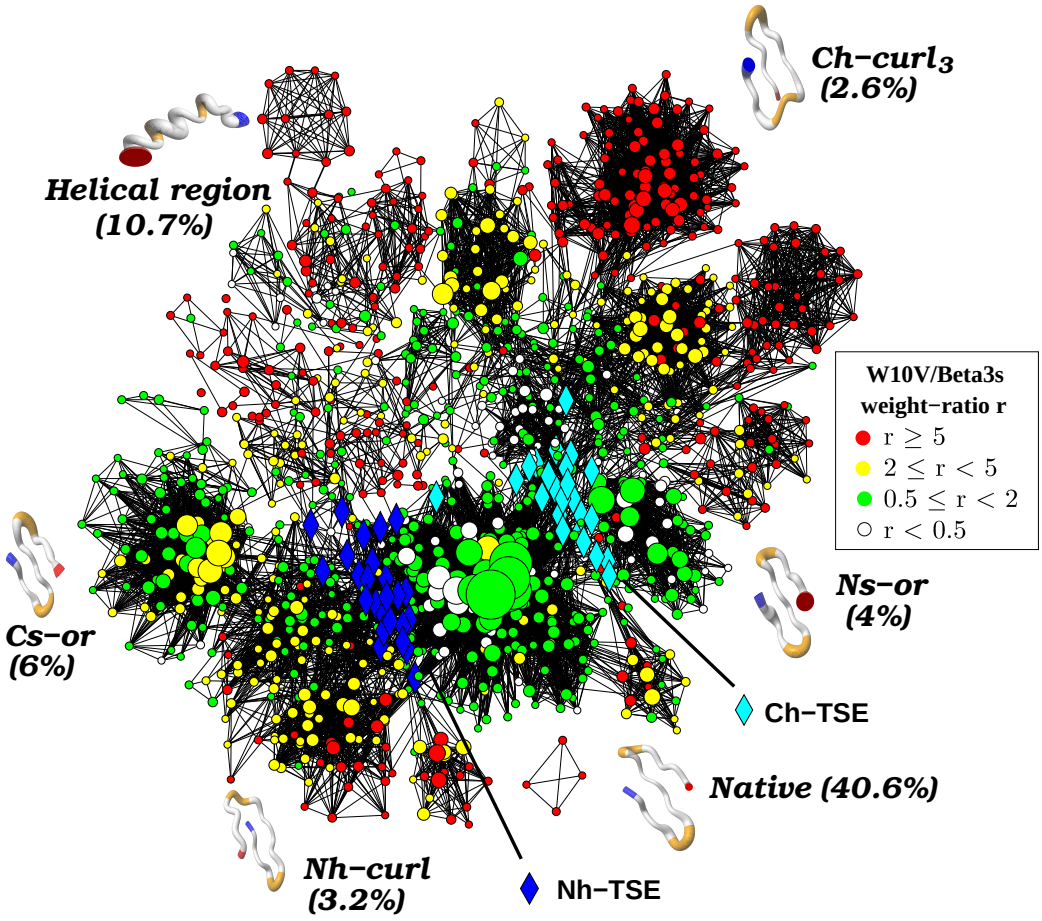


FIG. S1: The CSN of W10V. Each node (i.e., conformation) of the network represents a secondary structure string. The surface of each node is proportional to its statistical weight and only the 1192 nodes with at least 40 snapshots in W10V are shown to avoid overcrowding. Nodes are colored according to the W10V/Beta3s weight-ratio. Conformations in the most populated basins are shown by flexible tubes of variable diameter with N-terminus in blue, C-terminus in red and residues Gly6, Ser7, Gly14 and Ser15, which are at the two turns in the folded structure, in orange. The helical conformation shown on top left is the most populated helical string (`--IIHHHHHHHHHHHHHHH-`) in the W10V network. Blue and cyan diamonds emphasize TSE nodes with N-terminal and C-terminal hairpin formed, respectively. This Figure was made using visone (www.visone.de) and MOLMOL[1].

II. KINETIC GROUPING

During the 20 μs simulation time 120 and 105 folding events (i.e., visits to the native node) were observed for Beta3s and the W10V mutant, respectively, thus providing sufficient statistical sampling for the kinetic analysis. Unfolding events were defined as absence from the most populated (i.e., native) node longer than 10 ns.

To perform the kinetic all-against-all grouping, only significantly populated nodes with a statistical weight of at least 4×10^{-5}, i.e., 40 snapshots or more, have been employed (1430 for Beta3s, 1192 for W10V), which does not influence the kinetics since the original trajectory with all snapshots is considered for the p_{commit} calculations. 56% (59%) of the total weight in Beta3s (W10V) lies in nodes above the cutoff. This relatively low values are consistent with the weight distribution of nodes[2] which implies that most of the strings are very rare, occurring only once or twice in the simulation (also known as the *zero-frequency problem*[3,4]), and are not pronounced attractors. In fact, Beta3s and W10V spend 26%, respectively 24% of the time in nodes of weight one or two. Nodes with less than 40 snapshots are assigned to the basins identified by the heavy-node kinetic grouping in a post-processing step. Each "light" node is grouped to the basin to which the p_{commit} $\geq$0.5 criterion is fulfilled. If several candidates are possible, the most populated basin is chosen. The conformation space network (CSN) colored according to the kinetic grouping is illustrated in Fig. S2 and the most populated basins are listed in Table S-I.

A. Most populated strings in the native basin

The native basin includes 7569 and 5829 strings for Beta3s and W10V, respectively. The most populated are listed in Table S-II. Note that there is no correlation between the "geometrical" distance (i.e., number of different bits) and the kinetic distance from the native string. In fact, the strings in Table S-II have a geometrical distance of 4 to 8 and relax to the native string within 0.5 ns, whereas Ns-or (i.e., the string `-EEEESTTEEEEESSEEEE-`) has a geometrical distance of one and relaxes in 138 ns.

		Beta3s				W10V				
		Weight (%)		τ_f (ns)		Weight (%)		τ_f (ns)		
Conformation	Name	Node	Basin	Node	Basin	Node	Basin	Node	Basin	color
-EEEESSEEEEEESSEEEE-	native	5.59	36.42	–	–	8.76	40.63	–	–	green
Larger weight in Beta3s										
-EEEESTTEEEEESSEEE-	Ns-or	1.17	7.44	138	109	0.82	3.96	92	90	red
---SSGGG---EESSEETT-	Ch-curl$_1$	0.13	3.55	98	90	0	0	–	–	white ovals
---SSGGG-EESSTTTTEE-	Ch-curl$_2$	0.12	2.38	285	257	0	0	–	–	white circles
-----SS--EEEESSEEEE-		0.03	2.20	53	75	0.04	0.87	72	85	orange
-HHHHHHHHHHHS------	Helix$_{1-13}$	0.06	2.06	137	122	0.01	0.50	124	151	white squares
--EESSSEEEEEESSEEEE-		0.10	1.94	87	84	0.04	0.63	148	134	cyan
---SSGGG-EEESSSEEEE-		0.09	1.17	200	198	0	0	–	–	white rectangles
---SSSS--EEESTT-EEE-		0.06	0.94	316	263	0	0	–	–	white diamonds
Larger weight in W10V										
-EEEESSEEEEESSSEEEE-	Cs-or	0.26	3.56	63	70	0.65	6.02	69	75	olive
-EEEESSEEE---TT--B-	Nh-curl	0.04	0.49	59	58	0.13	3.23	69	69	blue
----STT---EEESSEEEE-		0.12	0.81	139	113	0.29	2.70	108	121	violet
--BSS-SSSEEE-STTEEE-	Ch-curl$_3$	0	0	–	–	0.12	2.58	104	105	white diamonds
--SSSSS--EEEESSEEEE-		0.03	0.81	103	97	0.09	2.09	111	100	yellow
-BSSSS---EEEESSEEEE-		0	0	–	–	0.02	0.28	61	53	white circles

TABLE S-I: Results of the kinetic grouping. Statistical weight of the native basin and the most populated free energy basins in the denatured state as identified by the kinetic grouping. The mean folding time (τ_f) to the native node are average values for snapshots in a node or basin. Conformations with names are shown by flexible tubes of variable diameter in Fig. S1 for W10V and in the main text for Beta3s: Ns-or, N-terminal strand out of register; Cs-or, C-terminal strand out of register; Nh-curl, curl-like conformation with structured N-terminal hairpin; Ch-curl, curl-like conformation with structured C-terminal hairpin. The colors indicated in the last column are those used in Fig. S2.

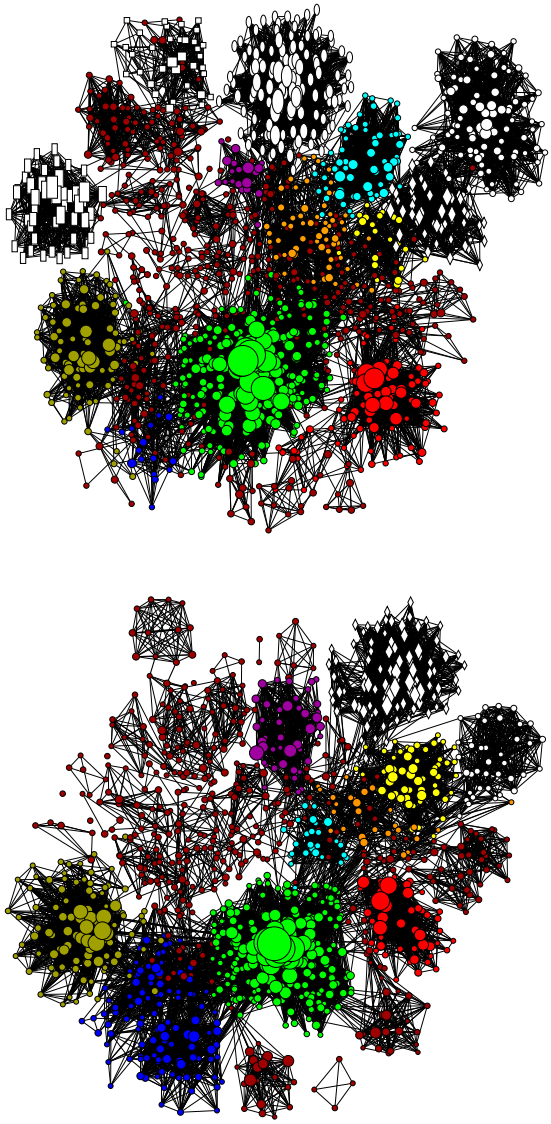


FIG. S2: The CSN of Beta3s (top) and W10V (bottom), colored according to the basins identified by the kinetic grouping. The coloring scheme is chosen such that basins with the corresponding most populated node have the same color in both peptides (last column in Table S-I). White is chosen if no significantly populated region exists in the conformational space of the other peptide. Nodes belonging to less populated basins and entropic regions are in brown.

Beta3s	%	τ_f (ns)	W10V	%	τ_f (ns)
-EEEESSEEEEEESSEEEE-	5.6	0	-EEEESSEEEEEESSEEEE-	8.8	0
-EEE-STTEEEEESSEEEE-	4.7	0.3	-EEE-STTEEEEESSEEEE-	5.2	0.3
-EEEESSEEEEE-STTEEE-	2.8	0.4	-EEEESSEEEEE-STTEEE-	4.9	0.2
-EEE-STTEEEE-STTEEE-	2.1	0.4	-EEE-STTEEEE-STTEEE-	2.9	0.3
-EEEESSEEEEEESSEEE--	1.9	0.4	-EEEESSEEEEESSTTEEE-	1.7	0.2
-EEESSTTEEEEESSEEEE-	1.4	0.3	-EEESSTTEEEEESSEEEE-	1.5	0.3
-EEE-TTTEEEEESSEEEE-	1.3	0.5	-EEE-STTEEEESSTTEEE-	1.0	0.3
-EEEESSEEEE-STTEE--	1.0	0.4	-EEESSTTEEEE-STTEEE-	0.9	0.2
-EEEESSEEEESSTTEEE-	0.9	0.5	-EEEESSEEEEEESSEEE--	0.9	0.5
-EEE-STTEEEEESSEEEE-	0.8	0.4	-EEE-TTTEEEEESSEEEE-	0.9	0.4

TABLE S-II: The ten most populated strings in the native basin with their relative populations and mean folding times (τ_f) for both peptides. Deviations from the native string are colored in red. Node-p_{fold} values[5] are larger than 0.98 for all nodes.

III. KINETIC PARTITIONING OF THE DENATURED STATE

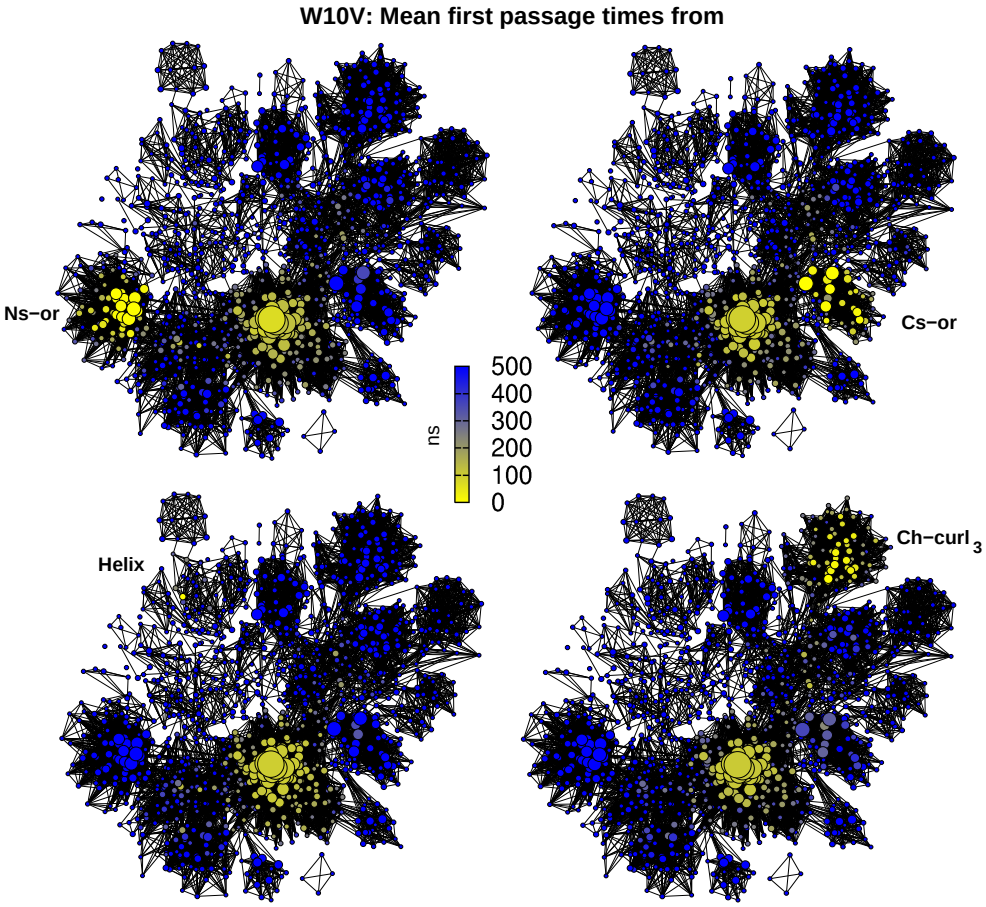


FIG. S3: The denatured state is kinetically partitioned. Mean first passage times from the most populated node of individual free energy basins in the unfolded state to all other nodes of the CSN of W10V are shown. Nodes within the basin of the starting node are visited relatively fast (yellow), indicating rapid intrabasin transitions and supporting the kinetic grouping analysis (Fig. S2 bottom). Equilibration between different unfolded basins (blue) is slower than reaching the folded state (olive) which shows that the denatured state is kinetically partitioned, i.e., no fast equilibration takes place between basins in the denatured state. In other words, the native state is a hub[2].

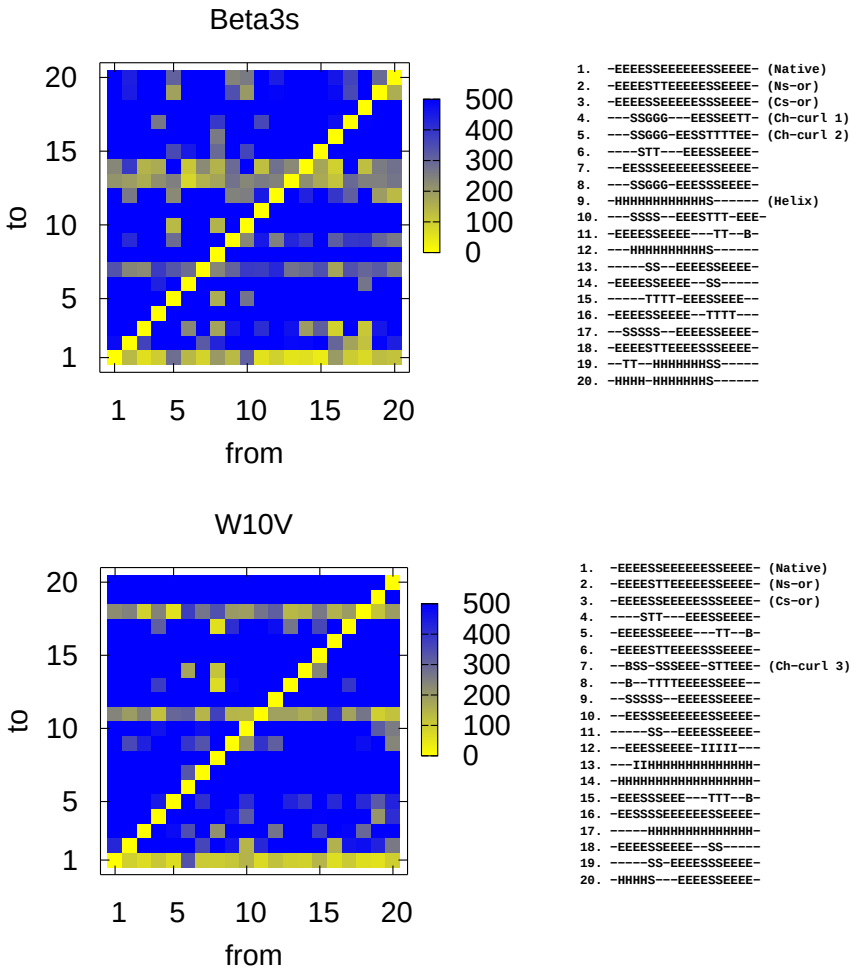


FIG. S4: Mean first passage times between the most populated nodes of the 20 most populated free energy basins as identified by kinetic grouping for Beta3s (top) and W10V (bottom). Transitions to the folded state (basin 1) are generally faster than transitions to other basins, indicating that the denatured state is partitioned by high barriers. Execptions are basin 13 and 14 (11 and 18) of Beta3s (W10V) that are involved in may folding events and turn out to lie on-pathway. Note that these basins are the same in both peptides and have either of the two hairpins fully formed, while the other hairpin is unstructured (-) except for the turns (SS at positions 6-7 or 14-15).

IV. NATIVE STRUCTURE IN THE DENATURED STATE

See Fig. S5.

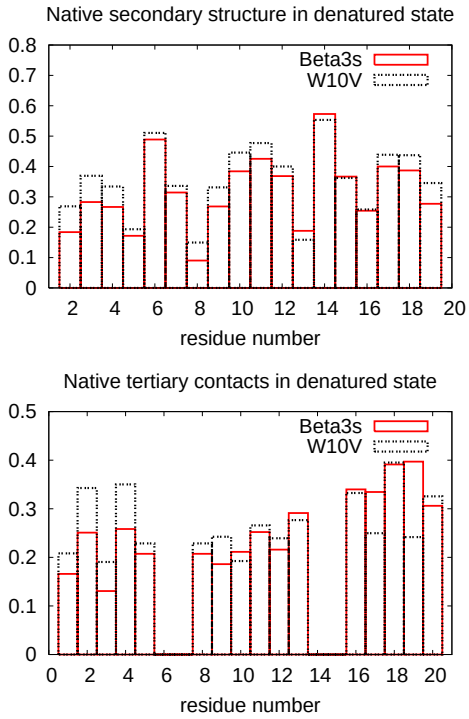


FIG. S5: The N-terminal hairpin (segment 2-11) has 19% higher content of native secondary structure (top) and 20% more native tertiary contacts (bottom) in the denatured state of the W10V mutant than the wild type peptide. Note the different y-axes.

V. DEPENDENCE OF THE KINETIC GROUPING ON τ_{commit}

A. The native basin

The isolation of the native basin is robust with respect to the commitment time which has been chosen as 1.6 ns (main text). Using 5 ns as τ_{commit} increases the population of the native basin only slightly from 36.4% to 38.5% in Beta3s and from 40.6% to 42.1% in W10V.

B. The denatured state

The value of τ_{commit} for the isolation of basins in the denatured state of Beta3s and its mutant has been set to 1 ns uniformly for all basins in order to calculate an all-against-all p_{commit}-matrix. The justification for this choice is that the relaxation times in important enthalpic basins lie within the order of magnitude of 1 ns, but transition from outside are two to three orders of magnitudes slower. The free energy profiles for the bottoms of the most populated enthalpic basins in Fig. S6 indeed show that they have only slightly different characteristic intra-basin relaxation times. The effect upon changing τ_{commit} from 0.5 ns to 5 ns has been investigated (Fig. S7). The analysis is robust for small time variations: no relevant changes in the isolation of basins is noticeable between 0.5 ns and 2 ns, which is the range where the most relevant basins have their maxima in Fig. S6. This means that the most relevant basins in the Beta3s unfolded state can be (at least approximatively) extracted using a constant commitment time. The same holds for W10V. Increasing τ_{commit} further, however, results in an almost trivial splitting of the CSN, where only the helical region is separated from the rest of the denatured state (red and white regions in Fig. S7, bottom right).

As mentioned in the Methods section of the main text, increasing values of τ_{commit} allow one to analyze different levels of ruggedness of the free energy surface. In fact, the Beta3s Cs-or and Nh-curl basins "merge" at τ_{commit}=2 ns (olive region in the bottom left part of Fig. S7), whereas Ns-or and Ch-curl$_{1,2}$ remain separated. This observation is consistent with the similar values of τ_f for Cs-or (70 ns) and Nh-curl (58 ns) and the different τ_f values of Ns-or (109 ns) and Ch-curl$_{1,2}$ (90 ns, 257 ns, Table S-I).

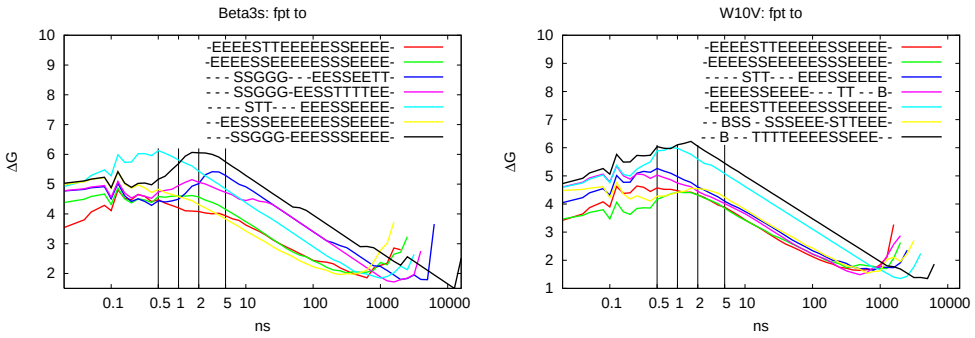


FIG. S6: Distribution of first passage times $P(fpt)$ to the most populated nodes of the largest basins in Beta3s (left) and W10V (right). Values are calculated as $\Delta G = -k_B T \ln(P(fpt))$ and plotted in kcal/mol using logarithmic binning without normalization of the bin size.

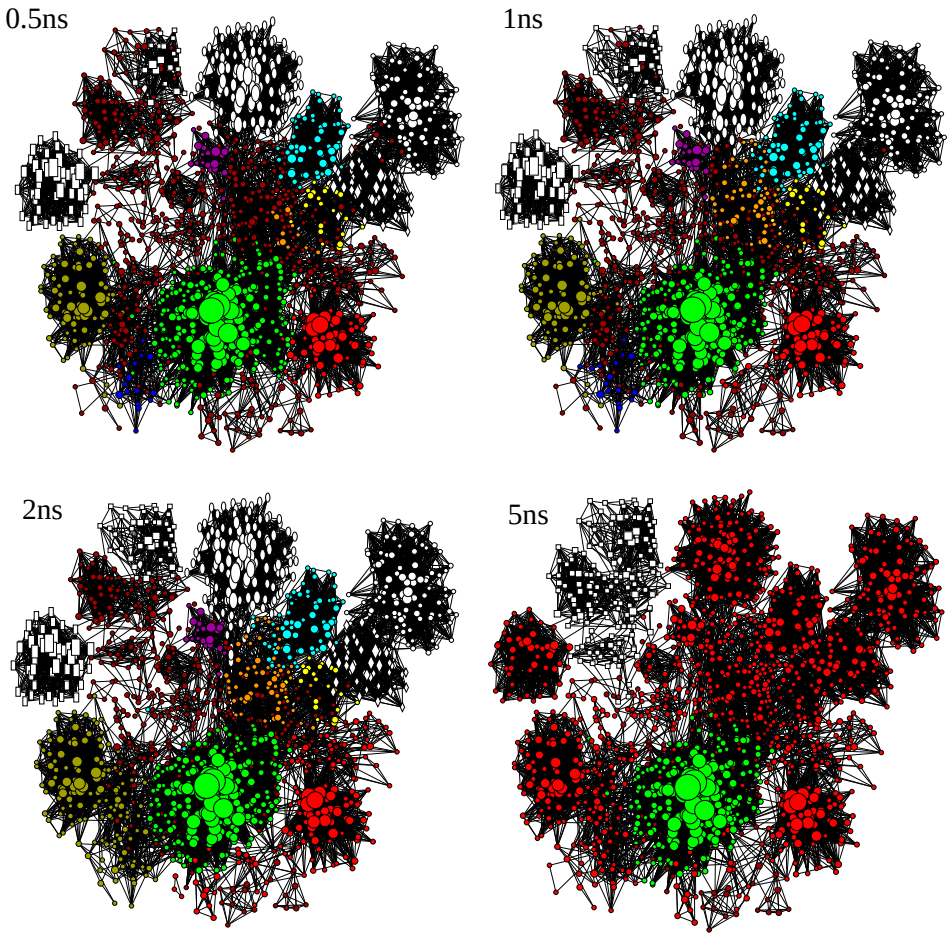


FIG. S7: Robustness of the kinetic grouping in the denatured state of Beta3s: slightly shorter (0.5 ns) and slightly longer (2 ns) τ_{commit} do not change the isolation of important basins compared to the value used in the main text (1 ns, see Table S-I for coloring scheme). The use of considerably longer times (5 ns), however, results in an almost trivial splitting of the CSN into one large region (red), where only the helical basin (white) is separated from the rest.

VI. THE TRANSITION STATE ENSEMBLE (TSE)

A. TSE nodes

See Table S-III.

Beta3s	node-p_{fold}	p_{fold} [a]	$\sigma_{p_{fold}}$ [b]	τ_f (ns)	W10V	node-p_{fold}	τ_f (ns)
-----GGG--EEESSEEE--	0.52	0.50	0.11	45	-----STT-EEEESSEEEE-	0.53	47
-EEEESSEEEES-GGG--B-	0.43	0.51	0.35	45	-EEEESSEEEE--TTT----	0.49	36
-----STT--EE-STTEE--	0.57	0.75	0.09	38	-EEEESSEEEE--HHHHHH-	0.46	33
----SSTT-EEEESSEEEE-	0.42	0.34	0.11	56	-EEEESSEEEE-SSTT----	0.53	38
-EEEESSEEEE-SSSS-EE-	0.59	0.54	0.23	48	-EEEESSEEEE--SSS----	0.45	37
-EEE-SSGGGEEESSEEEE-	0.41	0.34	0.09	9	--EEESSEEEESSSSEEE--	0.42	37
----BSSB--EESSTTEE--	0.51	0.68	0.16	51	-EEE-STTEEE--SSS-EE-	0.44	46
----STTT-EEEESSEEEE-	0.50	0.48	0.15	61	----SSTT-EEEESSEEEE-	0.44	72
-EEE-STTEEE--SSS----	0.41	0.26	0.28	42	-EEEESSEEEE-SSTT-TT-	0.51	18
-EE-SSSS-EEEESSEEE--	0.53	0.65	0.31	23	-----GGG-EEEESSEEEE-	0.47	61

TABLE S-III: The ten most populated TSE strings isolated by node-p_{fold} analysis. Green represents native secondary structure. Beta3s is more native in the C-terminal hairpin, while W10V shows native N-terminal hairpin predominance. Node-p_{fold}, p_{fold} and mean folding times (τ_f) are given. [a]p_{fold} was obtained by "shooting" 20 times from 10 individual snapshots to validate the above TSE conformations of Beta3s, and [b]$\sigma_{p_{fold}}$ is the standard deviation over the 10 snapshots. Note that τ_f's are in the order of half of the average folding time which is consistent with 50% unfolding events for trajectories passing through TSE nodes.

B. TSE robustness upon τ_{commit}

See Fig. S8.

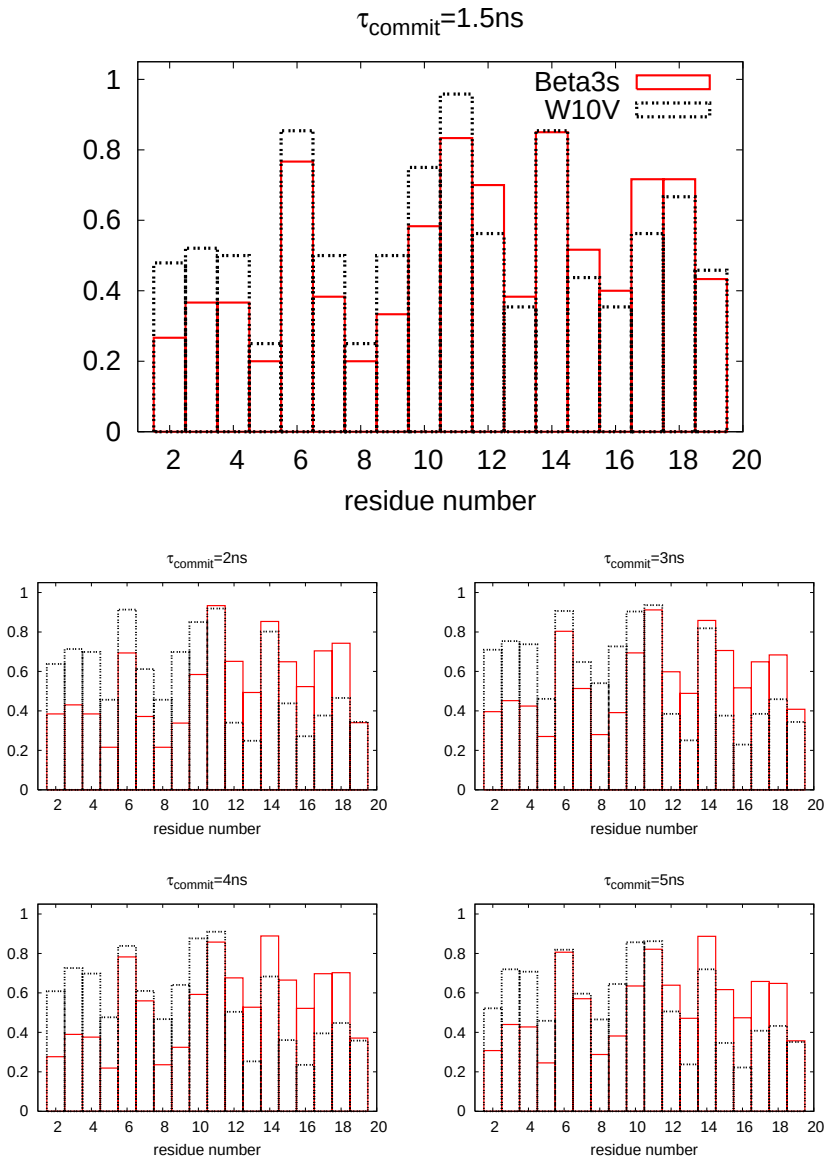


FIG. S8: Robustness of TSE selection upon τ_{commit}: the pathway switch remains evident for all choices between 0.5 ns and 5 ns. Only nodes with weight ≥ 20 have been considered.

C. Energetics of the TSE

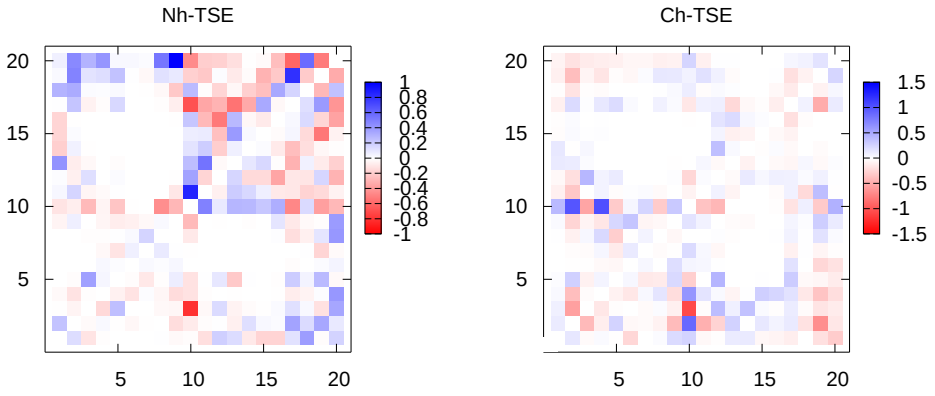


FIG. S9: The pairwise residue interaction energy differences (kcal/mol) between W10V and Beta3s for the Nh-TSE (left, N-hairpin predominant) and the Ch-TSE (right, C-hairpin predominant). The pairwise energy values in the native state are used as reference and subtracted from all values in the matrix. A red square indicates that the corresponding pair of residues has a more favorable interaction energy in the TSE of W10V than Beta3s. The upper and lower triangular matrices show the total and van der Waals energy, respectively, and their similarity indicates that most of the enthalpic effects originate from the difference in van der Waals energy.

D. Sampling of TSE nodes

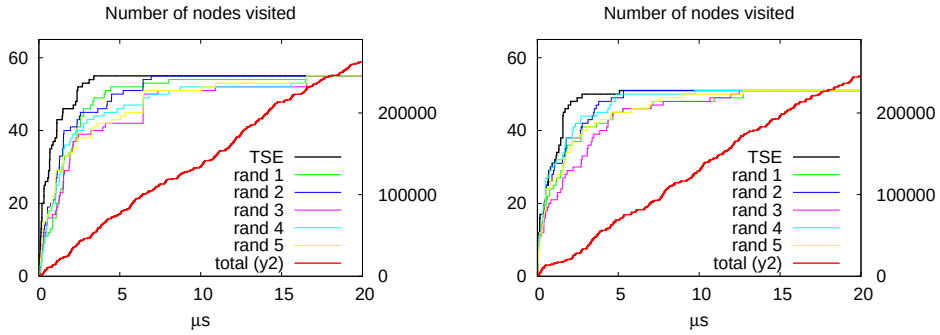


FIG. S10: The number of nodes visited depending on simulation time. The TSE of Beta3s (left) and W10V (right) are accessed completely after 3, respectively 5 μs. Random samples with the same size and weight distribution (weight between 20 and 110) as the two TSEs need in average more than 10 μs to visit all nodes at least once. This observation indicates that TSE nodes are involved in a large number of barrier-crossing events which implies that a node-p_{fold} value of 0.5 does not originate from only very few folding and unfolding events. The total number of nodes sampled during the simulation is shown in red (right axis in both plots). The difference in the evolution of the total number of nodes and the random samples comes from the fact that the latter contain only nodes with weight ≥ 20, while most of the contribution to the increase in total nodenumber comes from lower populated nodes.

VII. THE HELICAL ENSEMBLE

Interestingly, the distribution of helical content along the sequence shows that Beta3s is more helical than W10V in the central segment, i.e., residues 7-13 (Fig. S11). This observation is consistent with the fact that the side chain of valine has a destabilizing effect on the helical structure[6] because of the branching at the C_β carbon.

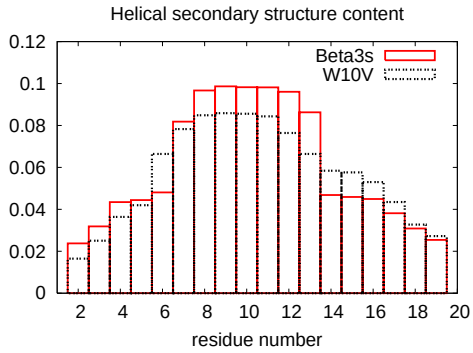


FIG. S11: The distribution of helical content along the sequence shows that Beta3s is more helical than W10V in the central segment.

The helical ensemble of W10V shows a slightly faster decay for the distribution of node weights (i.e., higher entropic character[7]) than the one of Beta3s (Fig. S12), which again reflects the destabilization of helical structure due to the valine side chain.

In order to assure that the distributions of node weights in the helical regions of the considered peptides are not affected by an undersampling problem, two tests have been carried out. Fig. S13 shows the helical weight distribution for the Beta3s simulation with a 50 times higher saving frequency nsavc (left), as well as the distribution of the first and second 10 μs against the full 20 μs simulation (right). In all cases the slope does not change significantly, thus a higher saving frequency and more sampling do not change the distribution, providing evidence that the entropic character of the helical region (i.e., the pronounced decay of the node-weight distribution) does not suffer from undersampling.

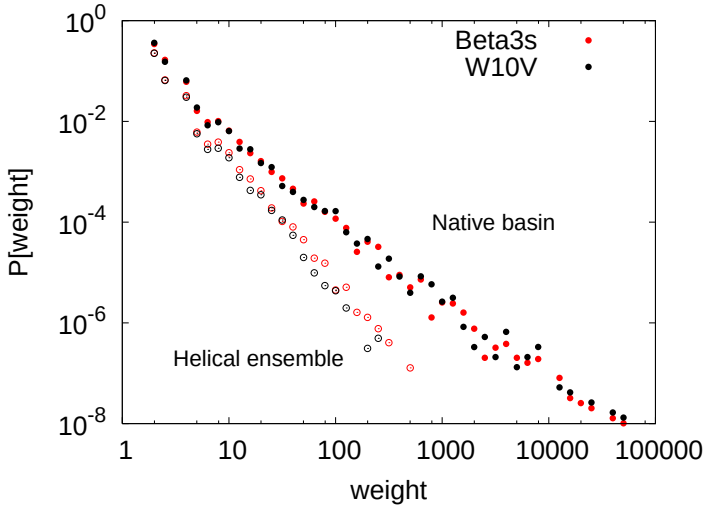


FIG. S12: Distribution of node weights. Logarithmic binning is used to reduce noise. The distributions of the enthalpic free energy basins Ns-or, Cs-or, Nh-curl, and Ch-curl show a very similar decay as the native basin and are not shown to avoid overcrowding.

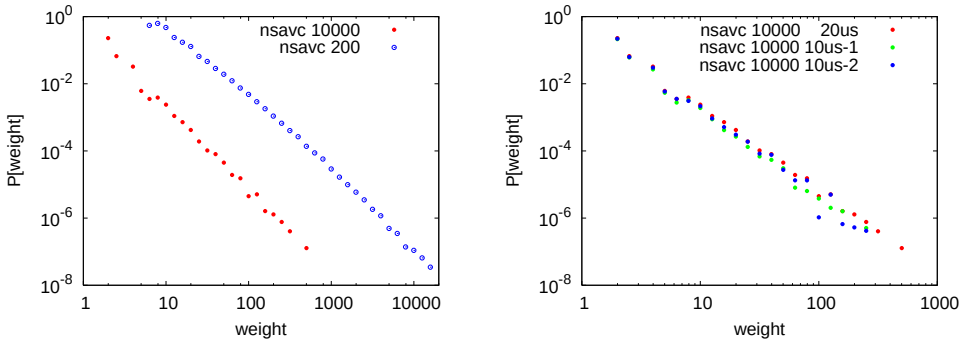


FIG. S13: Weight distribution of nodes in the helical basin. (left) A 50 times higher saving frequency (nsavc), as well as the splitting of the 20 μs simulation into two 10 μs-parts (right) do not change sensibly the slope of the distribution, indication that no undersampling problem exists in this case.

VIII. KINETIC GROUPING OF THE ALANINE DIPEPTIDE

Kinetic grouping is explained here using the alanine dipeptide which is a simple system yet containing the key features of a polypeptide chain. A total of 5×10^7 snapshots saved along a 1 μs MD trajectory at 300 K is used for this purpose. The main degrees of freedom are the dihedral angles ϕ and ψ. In the continuum solvent approximation used here[8] the projection of the free energy landscape onto ϕ and ψ shows four basins (see Fig. S14): C_{7eq}, α_R, C_{ax} and α_L. The most natural discretization of the phase space splits the (ϕ, ψ) space into cells. Using a 50×50 discretization of the Ramachandran map, 1821 nodes and 53995 links are visited during the 1 μs trajectory[7]. The most populated node in the system corresponds to the bottom of the C_{7eq} basin with coordinates ϕ=-86.4 and ψ=136.8.

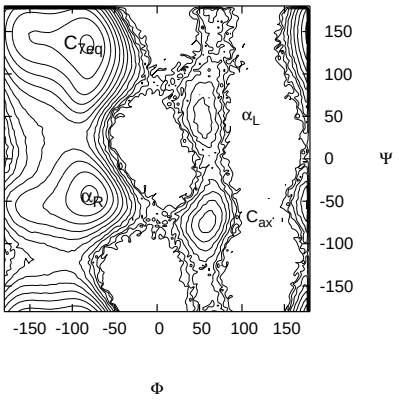


FIG. S14: The ϕ - ψ projection of the 1 μs MD simulation of the alanine dipeptide with the ACE2 implicit model of solvation[8]. Each contour line represents k_BT=0.6 kcal/mol.

Two procedures can be used for kinetic grouping. The first approach is a one-step all-against-all procedure which requires only one parameter determined by plotting fpt-distributions as in Fig. S6 and Fig. 6 of the main text for Beta3s and W10V. The second approach iteratively extracts basins (Fig. S17) by taking into account relaxation times of individual basins (Fig. S16).

A. Kinetic grouping: Simultaneous detection of basins

The simultaneous detection of basins is the procedure that was used for the denatured state of Beta3s and W10V (main text) because it is simpler and requires only a single τ_{commit} value. For all nodes populated above a certain cutoff, an all-against-all commitment probability (p_{commit}) matrix is calculated using a system-typical commitment time τ_{commit}. The weight-cutoff is introduced to make the analysis faster and robust, i.e., to avoid errors caused by nodes lying in high-energy regions. Two nodes are grouped if $p_{commit} \geq 0.5$, using the 80% criterion to reduce false negatives. The results for the alanine dipeptide nodes with $\tilde{w} \geq$300 are shown in Fig. S15. The four basins are identified correctly with $\tau_{commit} = 5$ ps which is the relaxation time to α_R (see Fig. S16). Notably, the nodes in the transition state region between C_{7eq} and α_R are not assigned to either of the two basins (black nodes). Using a commitment time significantly shorter (τ_{commit}=1.5 ps) or longer (τ_{commit}=10 ps) than 5 ps leads to a too detailed or too coarse split of the energy landscape, respectively (Fig. S15 top left and bottom), which shows how the choice of the commitment time is related to the allowed ruggedness of the surface.

B. Kinetic grouping: Iterative detection of basins

The existence of a single τ_{commit} for the simultaneous detection of all basins in more complicated systems is not necessarily guaranteed. A rigorous way for the isolation of basins by the use of kinetic information is to determine relaxation times for each region individually and then perform the kinetic grouping iteratively. The advantage of this approach is that heterogeneous relaxation times within a system are taken into account. Furthermore, there is no need to introduce a weight-cutoff to reduce the number of nodes as required in the simultaneous detection procedure. On the other hand, the analysis becomes more complicated and it is not clear how to automatize the choice of different values of the relaxation time. The procedure works as follows: in a first step, the distribution of the first passage times (fpts) to the most populated node is calculated. As indicated in Fig. S16 A, a value of τ_{commit}=10 ps is chosen to isolate the full C_{7eq} attractor region using $p_{commit} \geq$0.5 with the 80%-criterion (red region in Fig. S17). In a second step, the procedure is repeated for the most populated node that has not been grouped to the first basin. This node has coordinates

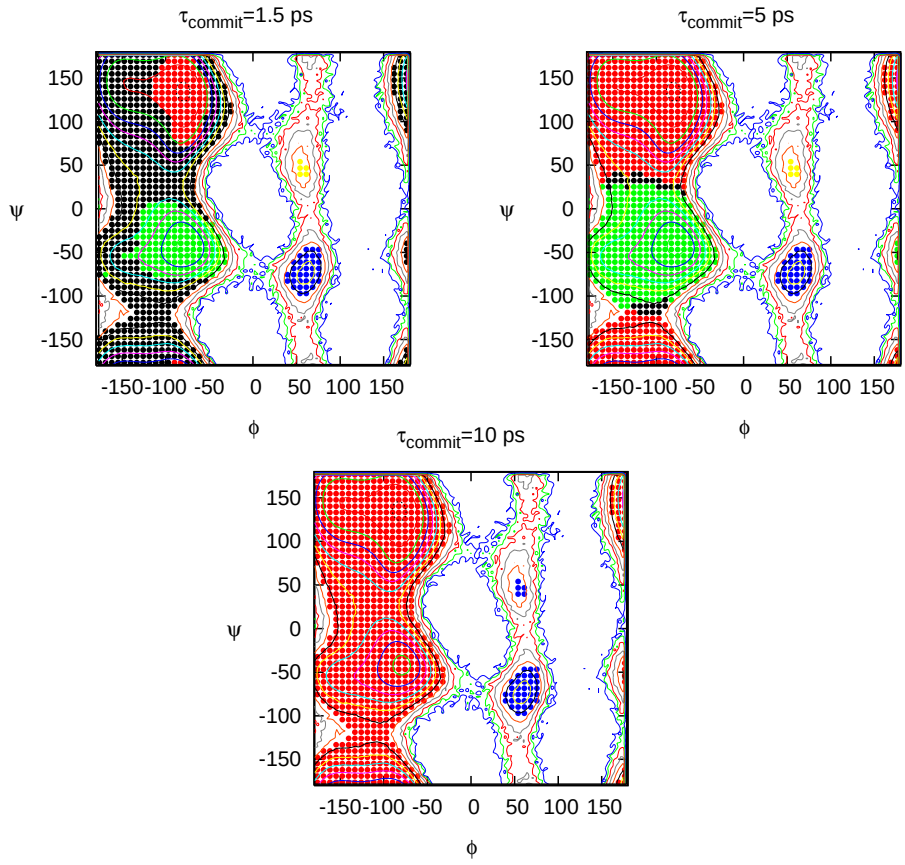


FIG. S15: Results of the simultaneous detection approach to isolate the basins in the alanine dipeptide using a single τ_{commit} value and only nodes with $\tilde{w}$ ≥300. Red, green, blue and yellow mark the region containing the bottom of the C_{7eq}, α_R, C_{ax} and α_L basin, respectively. Black dots are used for the remaining nodes which are split into several small groups for τ_{commit}=1.5 ps and 5 ps. The partition into four basins obtained using τ_{commit}=5 ps is very similar to the result of the iterative kinetic grouping (Fig. S17). On the other hand, using τ_{commit}=10 ps, the C_{7eq} and α_R basins as well as the C_{ax} and α_L basins are merged (red and blue regions).

ϕ=-79.2 and ψ=-43.2 and represents the bottom of the α_R region. The commitment time to identify the corresponding attractor region is 5 ps (Fig. S16 B) and a set of nodes is isolated (green region in Fig. S17) that has a marginal overlap with the C_{7eq} region. This intersection

contains putative transition state nodes. The procedure can be continued iteratively with τ_{commit}=1.5 ps for C_{ax} and α_L (Fig. S16 C and D), which leads to the splitting as indicated in Fig. S17. Interestingly, the regions isolated by the kinetic grouping analysis using either procedure (simultaneous or iterative detection) are comparable to those found by Markovian clustering[7] with granularity parameter $p = 1.2$.

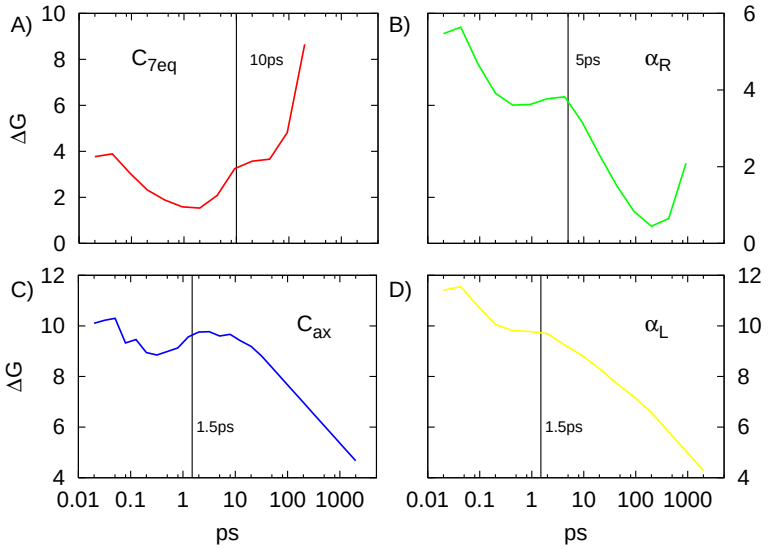


FIG. S16: Free energy profile as a function of temporal distance from the most populated node of (A) C_{7eq}, (B) α_R, (C) C_{ax} and (D) α_L. ΔG is calculated as $-k_BT \cdot ln(P(fpt))$ and plotted in kcal/mol. The fpt can be considered as a geometrically unbiased reaction coordinate. This projection is very useful to determine the transition between intra- and inter-basin relaxation, which is emphasized by the logarithmic binning without normalization of the bin size.

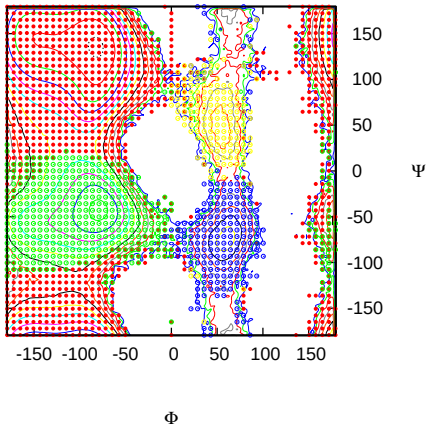


FIG. S17: Isolation of the free energy basins with the iterative kinetic grouping analysis. Coloring is consistent with Fig. S16. Black nodes are due to overlapping attractor regions of basins and occur close to transition state regions. Each contour line represents k_BT=0.6 kcal/mol.

Chapter 3

α-helix folding in the presence of structural constraints

[*PNAS*, **2008**, *105*,9588-9593]

α-Helix folding in the presence of structural constraints

Janne A. Ihalainen*†, Beatrice Paoli†‡, Stefanie Muff‡, Ellen H. G. Backus*, Jens Bredenbeck*, G. Andrew Woolley§, Amedeo Caflisch‡¶, and Peter Hamm*¶

*Physikalisch-Chemisches Institut and ‡Biochemisches Institut, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland; and §Department of Chemistry, University of Toronto, 80 Saint George Street, Toronto, ON, Canada M5S 3H6

Edited by William A. Eaton, National Institutes of Health, Bethesda, MD, and approved April 23, 2008 (received for review December 22, 2007)

We have investigated the site-specific folding kinetics of a photoswitchable cross-linked α-helical peptide by using single $^{13}C = {}^{18}O$ isotope labeling together with time-resolved IR spectroscopy. We observe that the folding times differ from site to site by a factor of eight at low temperatures (6°C), whereas at high temperatures (45°C), the spread is considerably smaller. The trivial sum of the site signals coincides with the overall folding signal of the unlabeled peptide, and different sites fold in a noncooperative manner. Moreover, one of the sites exhibits a decrease of hydrogen bonding upon folding, implying that the unfolded state at low temperature is not unstructured. Molecular dynamics simulations at low temperature reveal a stretched-exponential behavior which originates from parallel folding routes that start from a kinetically partitioned unfolded ensemble. Different metastable structures (i.e., traps) in the unfolded ensemble have a different ratio of loop and helical content. Control simulations of the peptide at high temperature, as well as without the cross-linker at low temperature, show faster and simpler (i.e., single-exponential) folding kinetics. The experimental and simulation results together provide strong evidence that the rate-limiting step in formation of a structurally constrained α-helix is the escape from heterogeneous traps rather than the nucleation rate. This conclusion has important implications for an α-helical segment within a protein, rather than an isolated α-helix, because the cross-linker is a structural constraint similar to those present during the folding of a globular protein.

cooperativity | infrared spectroscopy | molecular dynamics simulation | peptide folding

In many biomolecular systems, large changes can take place in response to a relatively small perturbation in the environment such as a variation in temperature, denaturant concentration, or the partial pressure of certain gases. Such an "all or nothing" phenomenon is termed a cooperative process. The classical example of cooperativity is the binding affinity of oxygen molecules to the four hemes of hemoglobin (1), which is a factor 100 to 1,000 times larger for the fourth oxygen molecule compared with the first. This leads to a sigmoidal dependence of oxygen binding on oxygen partial pressure with a sharp transition in a relatively small range of the latter. The folding of α-helices, which constitute one of the predominant secondary structures in many proteins, is often described in a similar manner (2): Once an entropically expensive nucleation process has occurred, i.e., a first helical turn with a hydrogen bond is formed, the zipping of additional hydrogen bonds is more likely because it is enthalpically favorable. A thermodynamic (statistical) treatment of the process leads to so-called nucleation-propagation-models (or zipper models), initially introduced by Zimm and Bragg (3) and Lifson and Roig (4). A large number of thermodynamic studies on α-helical peptides has been treated extremely successfully in terms of these models (5–8).

Cooperativity implies that the free energies of the two states of a system are balanced, however, in a way that enthalpic (ΔH) and entropic ($-T\ \Delta S$) contributions are large and compete against each other, leading to a characteristic sigmoidal transition as a function of the external control parameter. Because, in general, enthalpy and entropy vary in a nonsynchronous way as a function of some order parameter, the resulting free-energy surface $\Delta G \equiv \Delta H - T\Delta S$ will be uneven and in most cases will have a pronounced barrier. This is why this definition of cooperativity, which is based on thermodynamic arguments, often also has consequences for the kinetics of the transition between the two states. For example, in the classic case of oxygen binding to hemoglobin, a two-state allosteric model (i.e., the MWC model) can also explain the binding rate that increases with the number of already bound oxygen molecules (1). Cooperativity in protein folding reflects a two-state conformational distribution; its investigation requires a rigorous analysis of the folding transition (9). In the case of the helix–coil transition, high cooperativity would imply that once the rate-limiting nucleation step has occurred somewhere in the sequence, all subsequent helical turns would form at essentially the same time. Then, one common rate would be expected for all sites, corresponding to the nucleation rate (the propagation rate would not be detectable because it is very fast). Indeed, temperature-jump experiments on helix folding have successfully been described by "kinetic-zipper" models (10, 11), in which thermodynamic states of a nucleation-propagation model are linked by rate constants. However, because the cooperativity of isolated α-helices is weak, in particular when they are short, they do not fold in a two-state fashion, but rather with biexponential kinetics resulting from the coupling between nucleation and the only slightly faster diffusive elongation (10, 12).

It has recently been argued that even proteins that appear to be two-state folders can in fact be much more complex when reporting the folding free-energy surface on the level of individual protons by using NMR chemical-shift spectroscopy (13). IR spectroscopy together with site-selective isotope labeling offers the time resolution to perform site-selective folding studies also in a kinetic sense, even for the much faster folding of secondary structure motifs. The amide I′ (C = O stretch) vibrational mode is very sensitive to hydrogen bonding and dipole–dipole coupling among different peptide units (14), and isotope labeling of the carbonyl groups allows one to spectrally single out individual amino acids (12, 15). Gai and coworkers (12) demonstrated by $^{13}C = {}^{16}O$ labeling of groups of four

Author contributions: G.A.W., A.C., and P.H. designed research; J.A.I., E.H.G.B., and J.B. performed experiments; B.P. and S.M. performed simulations; G.A.W. contributed new reagents/analytic tools; J.A.I., B.P., S.M., E.H.G.B., J.B., A.C., and P.H. analyzed data; and J.A.I., A.C., and P.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

†J.A.I. and B.P. contributed equally to this work.

¶To whom correspondence may be addressed. E-mail: p.hamm@pci.uzh.ch or caflisch@bioc.uzh.ch.

This article contains supporting information online at www.pnas.org/cgi/content/full/0712099105/DCSupplemental.

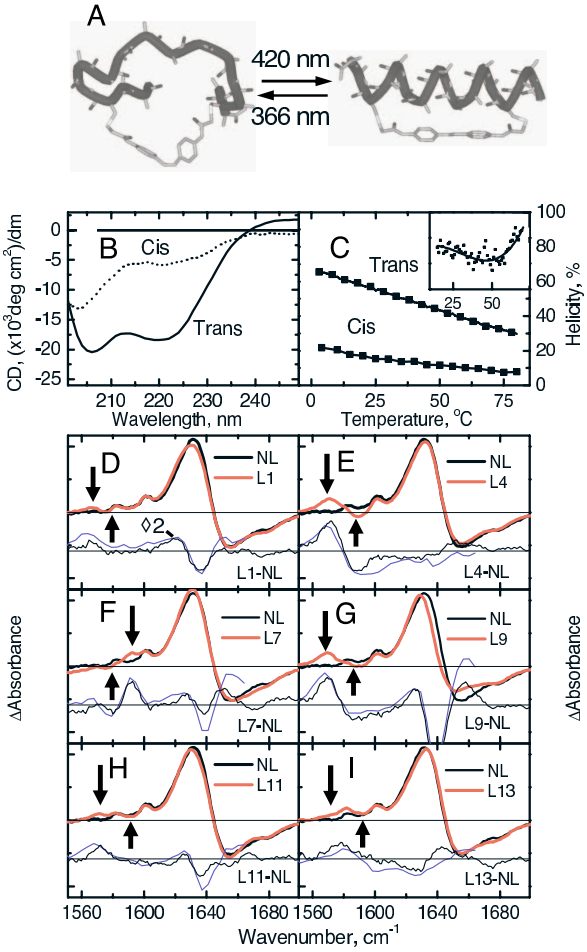


Fig. 1. Photoswitchable peptide and its steady-state spectra at room temperature. (*A*) Schematic drawing of the photoswitchable peptide in its *cis* (*Left*) and *trans* (*Right*) conformations. (*B*) CD-spectra of the peptide in its complete *trans* and *cis* states at room temperature. (*C*) The helicities in the *cis* and *trans* conformations at various temperatures. The *trans* spectrum was obtained in darkness, and the *cis*-spectrum was obtained under 365-nm illumination together with an estimate that 75% of the molecules are in the *cis* conformation (based on the UV-vis absorbance difference). (*Inset*) The first derivative of the *trans* CD data reveals the inflection point. (*D–I*) FTIR-difference spectra between the *trans* (under 436-nm illumination) and *cis* state (under 365-nm illumination) of each peptide at room temperature. Shown are the difference between not-labeled (NL) and labeled (L*X*, labeled at residue *X*) samples taken from the FTIR-spectra (thin black line) and from the late delay-time pump-probe spectra 30 μs after photoswitching (thin blue line). The arrows point to the frequency position from which the site-specific kinetics has been extracted.

subsequent alanines that the relaxation rates vary by ≈10% throughout the sequence. This result points to a more complex behavior of α-helix folding. Nevertheless, such a small degree of heterogeneity could be still modeled in the framework of nucleation-propagation models, taking into account position-dependent parameters for the different amino acids in the heteropolymer (16). However, as we will show, labeling of groups of several subsequent amino acids might still be too coarse-grained. If the kinetics in various sections of the peptide differ, one might expect variations even on the single amino acid level. Indeed, by using $^{13}C = {}^{18}O$ double labeling of a single amino acid in a small helix bundle protein, Dyer and coworkers (17) obtained considerably different melting curves and considerably different complex folding kinetics for that particular site, compared with the averaged signal.

As an alternative approach to study α-helix folding, we recently started to employ an azobenzene moiety as a photoswitchable structural constraint (see Fig. 1*A*) (18–20). Two cysteines are cross-linked in such a way that the azo-moiety in the *trans* (*cis*) conformation stabilizes (destabilizes) the helix, as deduced from CD spectroscopy (Fig. 1 *B* and *C*). This allows one to monitor both the folding and the unfolding direction of one and the same molecule at identical temperatures. We observed stretched exponential kinetics in both directions (20), in disagreement with the prediction from nucleation-propagation models that would reveal compressed exponential kinetics in the folding direction (21) (in fact, compressed exponential response has already been obtained in figure 10 of ref. 10, however, remained undiscussed). We therefore speculated in our previous work that the rate-limiting step in folding of our model systems is the escape from misfolded traps, which are consistently obtained in full-atom molecular dynamic (MD) simulations of small peptides (22–29).

Here, we present a comprehensive set of kinetic data of the folding of a photoswitchable α-helix (Ac-AACAK^{5}AAAAK^{10}AAACK^{15}A-NH_2) on the single amino acid level by employing $^{13}C = {}^{18}O$ labeling of an alanine-rich peptide, revealing strong variations in the folding kinetics and their temperature dependence along the sequence. The interpretation of the experimental data are corroborated by multiple implicit solvent (30) MD simulations (31) of folding of a similar cross-linked peptide (i.e., Ac-AACAR^{5}AAAAR^{10}AAACR^{15}A-NH_2) at low and high temperatures, and their network analysis (22, 29). Nonequilibrium MD simulations of an eight-residue cyclized peptide immersed in dimethyl sulfoxide have been published recently (32, 33) but those studies focused on the contributions to the frequency shift and did not mention pathways and kinetics of (helical) folding. The atomistic detail of the MD simulations and the large sampling of folding events (100 MD runs for a total of 0.4 ms) allow us to directly extract the conformational distribution from the trajectories. The comparison of experimental data and simulation results shows qualitatively similar folding kinetics and temperature dependence. Our analysis provides strong evidence that (*i*) the azo-cross-linker stabilizes kinetic traps originating from nonnative contacts that are nonexistant in isolated helices and that (*ii*) the azo-cross-linker finally destroys the already weak cooperativity of isolated α-helices. The analysis of the photoswitching simulations provides insights, at the atomic level of detail, on the folding mechanism and thereby explains the kinetic traces measured experimentally.

Results and Discussion

Noncooperative Helix Folding. Fig. 1*C* shows melting curves deduced from the fractional helicity at 222 nm [assuming $[\Theta]_{222} = -32{,}000$ deg cm^2/dm for the ellipticity of a fully folded helix (34)]. Within the limited temperature range, the sigmoidal dependence of the folding transition is hardly visible (see Fig. 1*C* *Inset*). Nevertheless, the helicity in the *trans* conformation varies slightly more than the relative temperature change ($\Delta T/T$ = 24%), hence we conclude that some small degree of cooperativity still remains, in the sense that folding enthalpy and folding entropy compete against each other. A comparison with melting curves from similar but nonlinked peptides (34, 35), which are steeper, indicates that the linker reduces the conformational flexibility of the unfolded ensemble and, hence, diminishes the entropic penalty of folding. It should be noted that the folding rates we observe for the azo-linked peptides (19, 20) are similar to those found in temperature-jump experiments of comparable unlinked peptides (10, 12, 36, 37) (a one-to-one comparison is

not possible because the sequences used in temperature-jump experiments are typically a bit longer) and that the amino acid sequence determines the folding rate to a significant extent. That is, at room temperature, the folding rate is 1,200 ns for Ac-EACAR^{5}EAAAR^{10}EAACR^{15}Q-NH$_2$ (19), 700 ns for Ac-AACAR^{5}AAAAR^{10}AAACR^{15}A-NH$_2$ (20), 600 ns for Ac-AACAK^{5}AAAAK^{10}AAACK^{15}A-NH$_2$ (this study), and 2,200 ns for Ac-EMCAR^{5}EMAAR^{10}EMACR^{15}Q-NH2 (data not shown). This variability indicates that interactions among the amino acid side chains and with the cross-linker may act as traps that are, indeed, rate determining.

Stationary FTIR-difference spectra between the two conformations of the photoswitchable helix are shown in Fig. 1 *D–I*. Upon folding, the unlabeled amide I′ band red-shifts with a positive signal ≈1633 cm^{-1} and a negative signal with additional substructure ≈1655 cm^{-1} and 1680 cm^{-1}, reporting on an overall strengthening of hydrogen bonding (14). Additional small bands are observed at lower frequencies that originate from both the ring modes of the azo-moiety (1,602 cm^{-1} and 1,580 cm^{-1}) (20) and from the isotope labels. By subtracting the FTIR-difference spectrum of the nonlabeled compound (termed NL throughout the text) from that of isotope-labeled compounds (termed L*X*, where *X* is the labeling position counted from the N terminus), the contributions of the latter can be isolated. As expected, they are downshifted by ≈65 cm^{-1} from the main band, exhibiting a dispersive shape with a sharp and distinct positive contribution in either case, whereas the negative contribution is, in general, broader and less clearly identified (see arrows in Fig. 1 *D–I*). The one notable exception is L7, which shifts to higher frequencies upon folding of the α-helix (Fig. 1*F*), opposite to all other residues. Because the frequency of the ^{13}C = ^{18}O vibration is related to the strength of hydrogen bonding, the only conclusion can be that hydrogen bonding of this particular site is stronger in the "unfolded" ensemble.

In the time-resolved experiments, helix folding is initiated by a subpicosecond laser pulse isomerizing the azo-photoswitch, and the formation of individual hydrogen bonds is detected in the amide I′ region as a function of time (19, 20). To single out the contribution from the isotope label, difference spectra between nonlabeled and labeled samples had to be taken. Furthermore, to suppress temperature-induced baseline effects, two kinetics were collected for each sample (Fig. 2 *A* and *B*; NL, open circles; L1, filled circles) at the frequency positions with the biggest deviations between nonlabeled and labeled sample (left and right from the ^{13}C = ^{18}O vibration, as indicated by arrows in Fig. 1). The result of subtracting the normalized signals from NL and L1 is shown in Fig. 2*C* (normalization was performed according to concentration obtained from the UV-vis spectrum and excitation power), and the difference between the traces at the two frequencies is shown in Fig. 2*D*. This "difference of difference signal" constitutes the site-selective folding trace and can be fitted (within signal-to-noise) with a single-exponential function shown as a solid line in Fig. 2*D* for L1. We note also that the difference–difference signals of the late time transient experiment agree with that from the FTIR-experiment (Fig. 1), validating the correctness of the background subtraction procedure. The site-selective folding traces labeled at all other positions are shown in Fig. 2 *E–J*. The immediate observation is that the rates scatter quite substantially without clear correlation between neighboring sites (Fig. 3*A*), thus, each peptide unit behaves independently, or noncooperatively. As a general trend, the signal is larger in the middle of the peptide, and sites 4, 7, and 9 reveal the most dominant signals (however, note again that L7 exhibits kinetics in the opposite direction; Fig. 2*F*). In line with other studies (e.g., ref. 38), a so-called "end-fraying effect" is observed as evidenced by smaller signals at the C and N termini. As a matter of fact, one would not expect any response from L13 and L16 at all, because they cannot form α-helical hydrogen bonds even in the folded conformation, and indeed, the signals

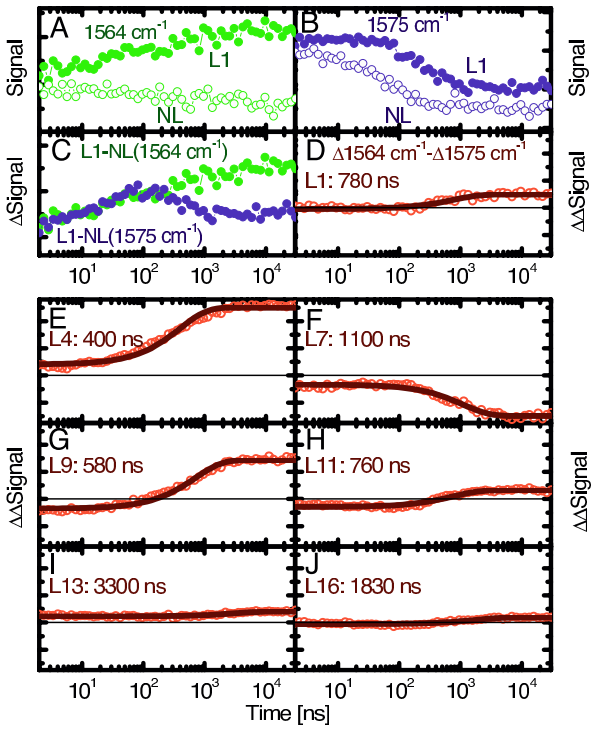


Fig. 2. Site-specific folding signals at 19°C. (*A* and *B*) Kinetic traces of the L1 and NL at 1,564 cm^{-1} (*A*) and at 1575 cm^{-1} (*B*). (*C*) The difference of the L1 and NL signals at these wavelengths. (*D*) The difference between the resulting signals together with its single exponential fit (solid line). (*E–J*) Site-selective folding signals of the other sites. The solid lines are single-exponential fits of the curves with the resulting time constant indicated.

are extremely small. Interestingly, the rates are correlated with the amplitudes of the corresponding signals (Fig. 3*A*); thus, the larger the driving force, the faster hydrogen bond formation.

Temperature Dependence. It is useful to plot the rates of folding of individual sites (i.e., the rates of helical hydrogen bond formation of individual backbone carbonyl groups) against reciprocal temperature, as measured by the isotopically labeled peptides (Fig. 3*C*). The individual sites exhibit considerably different temperature dependencies. The spread of rates is clearly bigger at low temperatures (a factor of approximately eight at 6°C), whereas they approach each other at higher temperatures (a factor of two at 45°C). Interestingly, summing up all site-signals with their relative intensities, the kinetics of the unlabeled band can be reproduced remarkably well (Fig. 3*E*). This result shows that our set of single-site labeled peptides represents the overall folding kinetics very well and, more importantly, that the averaged kinetics is just a trivial sum of the individual contributions. The average folding time of the NL peptide increases from approximately $\tau = 240$ ns to $\tau = 1{,}290$ ns when lowering the temperature from 45°C to 6°C and becomes more stretched with a stretching factor that decreases from $\beta = 1.00$ to $\beta = 0.71$ (fitting it with a function $\propto \exp[-(t/\tau)^{\beta}]$) (19, 20). The larger spread of rates reveals stretched exponential kinetics at low temperatures, whereas the more uniform values at higher temperatures result in a close to single-exponential response.

Qualitatively speaking, the MD simulations reveal very similar results. Just as in the experiment, the folding rates and amplitudes vary strongly from site to site (Fig. 3*B*), with a fraying effect toward the ends and a dip in the folding amplitudes in the middle

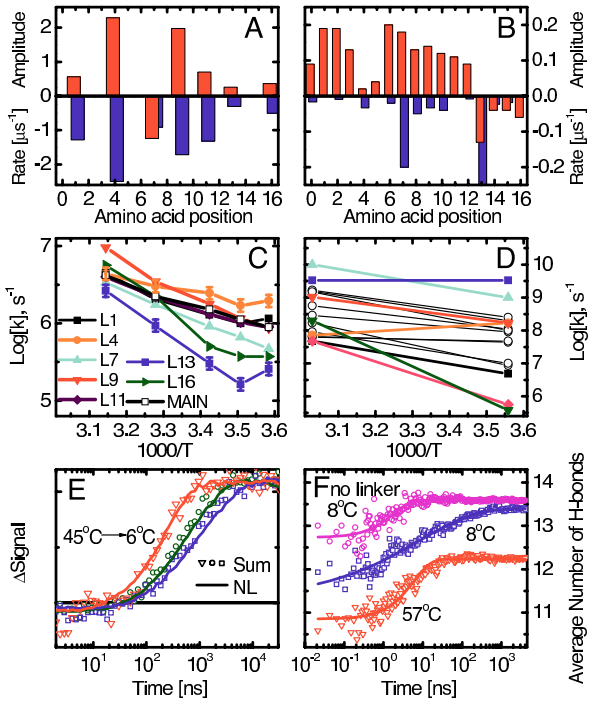


Fig. 3. Summary of the experimental (*A*, *C*, and *E*) and MD (*B*, *D*, and *F*) results. (*A* and *B*) Amplitudes (upward, red) and corresponding rates (downward, blue) at 19°C (*A*) and 8°C (*B*). Note that site 7 in the experiment, and sites 13–16 in the MD simulations, have inverted amplitudes. (*C* and *D*) Site-selective folding rates as a function of inverse temperature. In *D*, black circles are used for residues not measured experimentally. (*E*) Sum of all site signals (symbols) and amide I′ signal of the nonlabeled peptide (solid lines) at 6°C, 19°C, and 45°C. (*F*) Average number of hydrogen bonds along the MD simulations: 100 MD runs at 8°C with stretched exponential fit $h(t) = 13.6 - 1.6\exp(-t/29\ \text{ns})^{0.39}$ (blue), 50 MD runs at 57°C with single-exponential fit $h(t) = 12.2 - 1.4\exp(-t/6\ \text{ns})$ (red), and 50 MD runs of peptide without cross-linker at 8°C with single-exponential fit $h(t) = 13.6 - 0.9\exp(-t/2\ \text{ns})$ (magenta). The standard deviations are almost all <0.3 units. The experimental data in *E* are normalized, and the background has been removed, whereas the corresponding MD data in *F* are on their original scale.

of the helix (albeit not exactly at the same position). Also the temperature dependence observed in the MD simulations is in qualitative agreement with the experimental data (Fig. 3*D*), despite the larger spread of individual-residue rates, which is likely to originate mainly from the low friction coefficient [see supporting information (SI) *Text*]. Quantitative agreement is not expected because of the approximations inherent to the force field and implicit solvation model as well as the slightly different amino acid sequence. As an example, the fastest rate observed in MD (L13, blue square symbol in Fig. 3*D*) is the slowest in the experiment (Fig. 3*C*), which is in part a consequence of the very small amplitude. Nevertheless, essentially the same overall folding kinetics and temperature dependence emerge from the analysis of the MD simulations in terms of total number of hydrogen bonds involving carbonyl groups (Fig. 3*F*), which is the MD observable closest to the experimental signal. The folding signal from the MD runs at low temperature shows complex kinetics, clearly deviating from single-exponential behavior.

There are indications for deviations from single-exponential response even on the single-site level in both the experimental (Fig. 2) and the MD data (Fig. S3 and Tables S2 and S3), however, because of limited signal-to-noise, we do not discuss them in detail. Nevertheless, the largest contribution to the nonexponentiality of the overall signal originates from the spread of rates rather than from the nonexponentiality of the individual signals. Moreover, just as in the experiment, faster folding and single-exponential behavior are observed in the MD runs at high temperature (Fig. 3*F*, red). In particular, the ratio of folding times at low vs. high temperature is close: 29:6 in the simulations of the cross-linked peptide and 1,290:240 as measured experimentally. Finally, we simulated folding of the peptide without the cross-linker (Fig. 3*F*, magenta, an "experiment" that can only be performed on the computer) and obtain single-exponential response even at low temperature.

Origin of Complex Kinetics Explained by Network Analysis of MD Trajectories. In the example of oxygen binding to hemoglobin, cooperativity implies that the first step is the rate-limiting step (1). If the same were true for the folding of our photoswitchable α-helix, one common rate would be expected for all sites corresponding to the rate-limiting nucleation step. Such a common rate is not observed; instead, each amino acid site responds individually, reporting on various escape rates from different partially folded or trap states. The averaged signal obtained from the main band is then a nonspecific sum of these rates, and hence, the folding of the cross-linked α-helix is definitely not cooperative in the kinetic sense, although, from thermodynamic considerations (Fig. 1*C*), one might still deduce some degree of cooperativity. The complexity of the folding process is masked in an averaged signal, as suggested by previous atomistic simulation studies (22, 23), which have demonstrated that projecting the free energy on a single progress variable based on geometry, e.g., number of native contacts or rmsd from native, is not consistent with the complexity of the actual free-energy surface (25). Note that these projections (histogram-based free-energy profiles in Fig. S4) do not reveal any rate-limiting barrier for the photoswitchable α-helix.

The agreement of the MD results with the experiment justifies drawing detailed conclusions from the former. To that end, the folded (i.e., fully α-helical) state, and the most populated metastable states, can be isolated by grouping conformations according to fast relaxation along the MD trajectories, a procedure called kinetic grouping analysis (29). The advantage of this procedure with respect to a simple projection onto one or two progress variables is that structures (i.e., coordinate sets) are grouped into free-energy minima, not according to geometric characteristics, but rather according to the dynamics. The network analysis shows that escape from traps with a mixture of loop and helical content is rate limiting and that there are multiple parallel folding channels originating from a kinetically partitioned unfolded state (Fig. 4). The kinetic grouping analysis, using secondary structure strings (39), reveals that, of the four main folding channels, one starts from within the helical basin, whereas the remaining three start from the following metastable states: two C-terminal α-helical turns formed (Fig. 4*B*), only N-terminal α-helical turn formed (Fig. 4*C*), and only one C-terminal α-helical turn formed (Fig. 4*D*). The unfolded state is kinetically partitioned so that the traps are connected to the one folded state in a star-like manner, and the folded basin acts as a hub (22) for the interchange between misfolded or partially folded states. An example is shown in Fig. 4*D* where, starting from the trap with only the C-terminal turn formed, the peptide first reaches the fully α-helical state and then unfolds to a structure with broken N-terminal turn (as seen from the decreasing brightness of the yellow coloring of the nodes along this pathway). Importantly, different channels have barriers of different heights (Table S1), with rates ranging from ≈1/(10 ns) for folding from the helical basin (green in Fig. 4*A*) to ≈1/(1,000 ns) for folding from the structure with only N-terminal turn formed (cyan in Fig. 4*A*). Therefore, the spread observed in the site-specific rates originates from the heterogeneous degree of formation of individual helical hydrogen bonds in different traps.

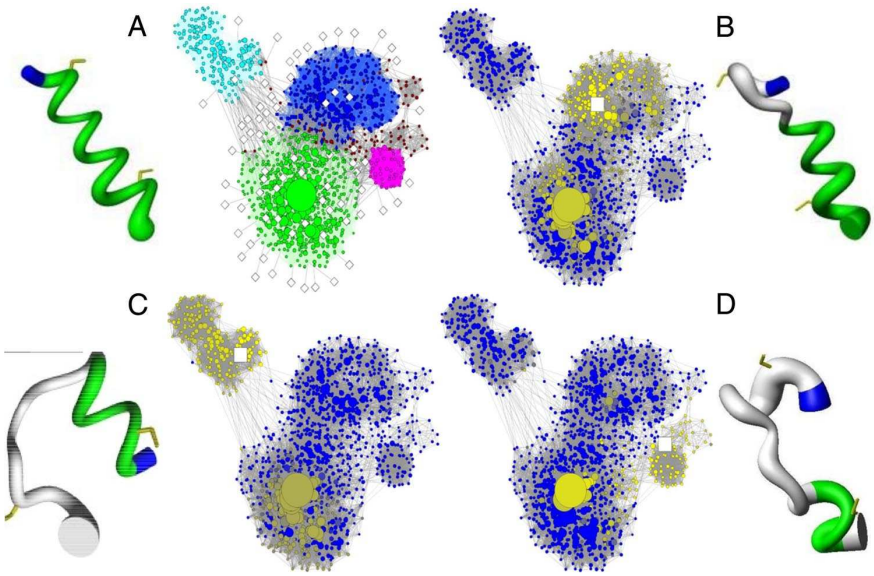


Fig. 4. The network analysis (22, 29) of the 100 MD runs at 281 K shows parallel folding channels. Each node (i.e., conformation) of the network represents a secondary structure string (39), and a link is a direct transition (within 20 ps) observed in the MD runs. The surface of each node is proportional to its statistical weight, and only the 1,387 nodes with at least 200 snapshots (≈96.7% of total sampling) are shown to avoid overcrowding. (*A*) The free-energy basins, i.e., native (green nodes) and metastable states [identified by kinetic grouping analysis (29) using a commitment time of 10 ns to group conformations that interconvert rapidly], are shown with different colors, and their characteristics are listed in Table S1. Within each basin, nodes and intrabasin links are shown with the same color, and interbasin links are colored in gray. White diamonds indicate the starting points of 82 of the 100 folding runs, whereas the remaining 18 runs reached directly the most populated node (i.e., fully formed α-helix, large green circle) and are not shown. An enlarged version of the network is shown in Fig. S2. (*B*–*D*) Nodes are colored according to the values of the mean first passage time (29) from the most populated node (white squares) of individual metastable states to all other nodes. The time scale goes from 0 (yellow) to >2 μs (blue). The coloring shows that the unfolded state is kinetically partitioned, and the folded (i.e., fully α-helical) state acts as a hub. Visits to unfolded metastable states different from the starting one require a much longer time than reaching the folded state. Representative structures of the folded state and each metastable state are shown by flexible tubes of variable diameter reflecting conformational disorder, with α-helical turns in green, loop segments in gray, N terminus in blue, and cysteine side chains in yellow for emphasizing the position of the linker.

The interpretation of the experimental data, thanks to the folding mechanism and pathways extracted from the MD simulations, provides strong evidence that the rate-limiting step of helix folding (at low temperature) is exactly the escape rate from a few metastable states (some of them stabilized by nonnative contacts) rather than the commonly assumed nucleation step.

Conclusion

The kinetics and mechanism of folding of a photoswitchable cross-linked α-helix have been investigated by a combined experimental–simulation study. At low temperatures, the hydrogen bond formation rates of different sites scatter by almost one order of magnitude, whereas they approach each other at higher temperatures. The spread of rates is significantly larger than the 10% effect observed for an isolated helix (12) and appears to be too large to be consistent with conventional nucleation-propagation models along the lines of ref. 16. Furthermore, with group L7, we directly observe a nonnative contact in the misfolded ensemble, the existence of which, however, is neglected in nucleation-propagation models. On the other hand, good agreement with an all-atom MD simulation is obtained, which justifies drawing interpretations at atomic resolution from the MD results. Notably, the MD analysis unmasks discrete traps (i.e., nonnative free-energy basins) along parallel folding pathways, which render the overall kinetics nonexponential. The cross-linker actually stabilizes these traps, as observed from the difference in response in MD runs with and without cross-linker (Fig. 3*F*). However, in contrast to previous studies, where noncooperative folding has been interpreted as barrierless folding (40, 41), we argue here that a completely different scenario, i.e., a few traps in the unfolded state that are separated from the native basin by barriers of different heights, may lead to a similar noncooperative behavior. This scenario (with barriers of different heights) is different from a barrierless landscape, but they share a higher population of conformations with intermediate compactness with respect to the two-state behavior (42).

Isolated α-helices fold in a marginally cooperative manner, as seen by the somewhat steeper melting curve (34, 35) than in Fig. 1*C* and the only small variation of folding rates along the peptide chain (12). If α-helix folding were cooperative, one could think of it as binary (all or nothing) when studying the folding of tertiary structures in larger proteins. However, addition of an azo-cross-linker as a switchable structural constraint finally destroys the already weak cooperativity of isolated α-helices. Structural constraints of this sort might also exist for a helical segment in a larger protein by the very fact that the helix is connected through the backbone to the rest of the polypeptide chain, and its flexibility is restrained by tertiary contacts with other pieces of the protein. As such, the folding of secondary and tertiary structures cannot be thought of as decoupled. Paradoxically, the folding of the cross-linked α-helix might be closer to the natural situation inside a globular protein than that of an isolated helix.

Materials and Methods

Experimental. Synthesis of the molecule, Ac-AACAK^{5}AAAAK^{10}AAACK^{15}A-NH_2, cross linked by a azo-moiety acting as a photoswitch and with $^{13}C = {}^{18}O$-labeled alanine at eight different positions, was performed as described (18). IR pump-probe spectroscopy with delays ranging from 10 ps to 40 μs was performed by using two electronically synchronized Ti:S laser systems, one of which was frequency doubled to generate pump pulses at 420 nm and the second pumped an IR-OPA to obtain broadband IR probe pulses (19, 20).

Computational. MD simulations were performed by using the CHARMM program package (31) using standard procedures and an implicit solvent (30). The force

field parameters for the azo-moiety were derived from the PARAM19 for the amide backbone and phenyl ring of Phe as well as from ref. 43 for the dihedral angles of the central N = N bond (Fig. S1). An equilibrium ensemble with the cross-linker in the *cis* conformation was sampled by two runs of replica exchange MD (44). After instantaneously switching the torsional potential of the central N = N bond to one that strongly favors the trans-configuration, ensembles of nonequilibrium Langevin dynamics runs of 4 μs each were started from the *cis* equilibrium ensemble at both 330 K and 281 K. Network analysis (22, 29) of the resulting nonequilibrium trajectories was supported by the program WORDOM (45). Because of the implicit solvent model, both the absolute temperatures and rates are somewhat arbitrary. In the comparison with the experiments, we therefore focus on trends, rather than the absolute values.

For a more detailed account of materials and methods, see *SI Text*.

ACKNOWLEDGMENTS. We thank Riccardo Pellarin for suggesting the control runs without cross-linker, Gianluca Interlandi for help in the initial set-up of the MD simulations, Rolf Pfister for the synthesis of the molecules, Francesco Rao for interesting discussions, Ben Schuler for instructive discussions and for the access to the HPLC-equipment, Bernhard Spingler and Philipp Antoni for the access to the CD equipment, and Jan Helbing and Andrea Prunotto for technical assistance. The MD simulations were run on the Matterhorn cluster of the University of Zürich. This work was supported by Swiss National Science Foundation Grants 200020-107492/1 (to P.H.) and 205320-118214 (to A.C.) and by a fellowship of the "Forschungskredit" of the University of Zürich (to E.H.G.B.).

1. Eaton WA, Henry ER, Hofrichter J, Mozzarelli A (1999) Is cooperative oxygen binding by hemoglobin really understood? *Nat Struct Biol* 6:351–358.
2. Creighton TE (1993) *Proteins*. (Freeman, New York).
3. Zimm BH, Bragg JK (1959) Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys* 31:526–535.
4. Lifson S, Roig A (1961) On the theory of helix–coil transition in polypeptides. *J Chem Phys* 34:1963–1974.
5. Scholtz JM, Qian H, York EJ, Stewart JM, Baldwin RL (1991) Parameters of helix–coil transition theory for alanine-based peptides of varying chain lengths in water. *Biopolymers* 31:1463–1470.
6. Muñoz V, Serrano L (1995) Helix design, prediction and stability. *Curr Opin Biotechnol* 6:382–386.
7. Doig AJ (2002) Recent advances in helix–coil theory. *Biophys Chem* 101: 281–293.
8. Scheraga HA, Vila JA, Ripoll DR (2002) Helix–coil transitions re-visited. *Biophys Chem* 101–102:255–265.
9. Kaya H, Chan HS (2000) Polymer principles of protein calorimetric two-state cooperativity. *Proteins* 40:637–661.
10. Thompson PA, Eaton WA, Hofrichter J (1997) Laser temperature jump study of the helix–coil kinetics of an alanine peptide interpreted with a "kinetic zipper" model. *Biochemistry* 36:9200–9210.
11. Thompson PA, *et al.* (2000) The helix–coil kinetics of a heteropeptide. *J Phys Chem B* 104:378–389.
12. Huang C-Y, *et al.* (2002) Helix formation via conformation diffusion search. *Proc Natl Acad Sci USA* 99:2788–2793.
13. Sadqi M, Fushman D, Muñoz V (2006) Atom-by-atom analysis of global downhill protein folding. *Nature* 442:317–321.
14. Krimm S, Bandekar J (1986) Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv Protein Chem* 38:181–364.
15. Silva RAGD, Kubelka J, Bour P, Decatur SM, Keiderling T (2002) A site-specific conformational determination in thermal unfolding studies of helical peptides using vibrational circular dichroism with isotopic substitution. *Proc Natl Acad Sci USA* 97:8318–8323.
16. Doshi U, Muñoz V (2004) The principles of α-helix formation: Explaining complex kinetics with nucleation-elongation theory. *J Phys Chem B* 108:8497–8506.
17. Brewer SH, Song B, Raleigh DP, Dyer RB (2007) Residue specific resolution of protein folding dynamics using isotope-edited infrared temperature jump spectroscopy. *Biochemistry* 46:3279–3285.
18. Kumita JR, Smart OS, Woolley GA (2000) Photo-control of helix content in a short peptide. *Proc Natl Acad Sci USA* 97:3803–3808.
19. Bredenbeck J, Helbing J, Kumita JR, Woolley GA, Hamm P (2005) α-Helix formation in a photoswitchable peptide tracked from picoseconds to microseconds by time resolved IR spectroscopy. *Proc Natl Acad Sci USA* 102:2379–2384.
20. Ihalainen JA, *et al.* (2007) Folding and unfolding of a photoswitchable peptide. *Proc Natl Acad Sci USA* 104:5383–5388.
21. Hamm P, Helbing J, Bredenbeck J (2006) Stretched versus compressed exponential kinetics in α-helix folding. *Chem Phys* 323:54–65.
22. Rao F, Caflisch A (2004) The protein folding network. *J Mol Biol* 342:299–306.
23. Krivov SV, Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc Natl Acad Sci USA* 101:14766–14770.
24. Hummer G, Garcia AE, Garde S (2001) Helix nucleation kinetics from molecular simulations in explicit solvent. *Proteins* 42:77–84.
25. Caflisch A (2006) Network and graph analyses of folding free energy surfaces. *Curr Opin Struct Biol* 16:71–78.
26. Chekmarev SF, Krivov SV, Karplus M (2006) Folding of ubiquitin: A simple model describes the strange kinetics. *J Phys Chem B* 110:8865–8869.
27. Chowdhury S, Zhang W, Wu C, Xiong G, Duan Y (2003) Breaking non-native hydrophobic clusters is the rate limiting step in the folding of an alanine-based peptide. *Biopolymers* 68:63–75.
28. Makowska J, *et al.* (2006) Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. *Proc Natl Acad Sci USA* 103:1744–1749.
29. Muff S, Caflisch A (2008) Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins Struct Funct Bioinf* 70:1185–1195.
30. Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins Struct Funct Bioinf* 46:24–33.
31. Brooks BR, *et al.* (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.
32. Nguyen PH, Stock G (2006) Nonequilibrium molecular dynamics simulation of a photoswitchable peptide. *Chem Phys* 323:36–44.
33. Nguyen PH, Gorbunov RD, Stock G (2006) Photoinduced conformational dynamics of a photoswitchable peptide: A nonequilibrium molecular dynamics simulation study. *Biophys J* 91:1224–1234.
34. Marqusee S, Robbins VH, Baldwin RL (1989) Unusually stable helix formation in short alaninen-based peptides. *Proc Natl Acad Sci USA* 86:5286–5290.
35. Huang C-Y, Klemke JW, Getahun Z, DeGrado WF, Gai F (2001) Temperature-dependent helix–coil transition of an alanine based peptide. *J Am Chem Soc* 123:9235–9238.
36. Werner JH, Dyer RB, Fesinmeyer RM, Andersen NH (2002) Dynamics of the primary processes of protein folding: Helix nucleation. *J Phys Chem B* 106:487–494.
37. Gooding EA, *et al.* (2005) The effects of individual amino acids on the fast folding dynamics of α-helical peptides. *Chem Commun* 5985–5987.
38. Rohl CA, Baldwin RL (1998) Deciphering rules of helix stability in peptides. *Methods Enzymol* 295:1–26.
39. Andersen CAF, Palmer AG, Brunak S, Rost B (2002) Continuum secondary structure captures protein flexibility. *Structure (London)* 10:174–184.
40. Garcia-Mira MM, Sadqi M, Fischer N, Sanchez-Ruiz JM, Muñoz V (2002) Experimental identification of downhill protein folding. *Science* 298:2191–2195.
41. Yang WY, Gruebele M (2003) Folding at the speed limit. *Nature* 423:193–197.
42. Knott M, Chan HS (2006) Criteria for downhill protein folding: Calorimetry, chevron plot, kinetic relaxation, and single-molecule radius of gyration in chain models with subdued degrees of cooperativity. *Proteins* 65:373–391.
43. Carstens, H (2004) Conformation dynamics of light-stable peptide: Molecular dynamics simulations and data-driven model-building. PhD thesis (Ludwig Maxmilians University, Munich).
44. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314:141–151.
45. Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) WORDOM: A program for efficient analysis of molecular dynamics simulations. *Bioinformatics* 23:2625–2627.

Chapter 4

One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: new insights into the folding process.

[*J. Phys. Chem. B*, **2008**, *112(29)*,8701-8714]

One-Dimensional Barrier-Preserving Free-Energy Projections of a β-sheet Miniprotein: New Insights into the Folding Process

Sergei V. Krivov,[†,‡] Stefanie Muff,[‡,§] Amedeo Caflisch,[*,§] and Martin Karplus[*,†,∥]

Laboratoire de Chimie Biophysique, ISIS F-67000, Strasbourg, France, Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland, and Department of Chemistry & Chemical Biology, Harvard University, Cambridge, Massachusetts 02138

Received: December 18, 2007; Revised Manuscript Received: March 12, 2008

The conformational space of a 20-residue three-stranded antiparallel β-sheet peptide (double hairpin) was sampled by equilibrium folding/unfolding molecular dynamics simulations for a total of 20 μs. The resulting one-dimensional free-energy profiles (FEPs) provide a detailed description of the free-energy basins and barriers for the folding reaction. The similarity of the FEPs obtained using the probability of folding before unfolding (p_{fold}) or the mean first passage time supports the robustness of the procedure. The folded state and the most populated free-energy basins in the denatured state are described by the one-dimensional FEPs, which avoid the overlap of states present in the usual one- or two-dimensional projections. Within the denatured state, a basin with fluctuating helical conformations and a heterogeneous entropic state are populated near the melting temperature at about 11% and 33%, respectively. Folding pathways from the helical basin or enthalpic traps (with only one of the two hairpins formed) reach the native state through the entropic state, which is on-pathway and is separated by a low barrier from the folded state. A simplified equilibrium kinetic network based on the FEPs shows the complexity of the folding reaction and indicates, as augmented by additional analyses, that the basins in the denatured state are connected primarily by the native state. The overall folding kinetics shows single-exponential behavior because barriers between the non-native basins and the folded state have similar heights.

Introduction

Protein and peptide folding from the very broad ensemble of denatured conformations to the well-defined native state is a very complex unimolecular reaction because of the many degrees of freedom of the system.[1] The conformational transitions involved, like those of other chemical reactions, are governed by the free-energy surface.[2] During the folding process, the loss of configurational entropy of the protein chain is approximately counterbalanced by the more favorable interactions among the protein atoms, modulated by the effect of the solvent. Thus, although the enthalpy and entropy of folding can be large at physiological temperatures, the free energy stabilizing the native state is normally only 10 kcal/mol or less, independent of the size of the protein.[3] The loss of configurational entropy during folding is thought to be primarily responsible for the experimental activation barrier, observed in many so-called two-state proteins.[4,5] Consequently, the major role played by the entropic contributions in protein folding, in contrast to most simple reactions,[6] requires an analysis of the free-energy surface; that is, knowledge of the potential energy surface is not sufficient.[6–10]

Although a protein has many degrees of freedom (with the φ and ψ backbone dihedral angles of the amino acids being particularly important), the common way to investigate the free-energy surface is to display it as a function of a small (usually only one to two) number of order parameters. A commonly used coordinate is the fraction of native contacts, Q.[11] Q appears to be a satisfactory approximate reaction coordinate for Go-model proteins[12,13] because favorable interactions occur only between residues in contact in the folded state.[14] On the other hand, for transferable potentials (e.g., those based on physicochemical principles, such as AMBER, CHARMM, and OPLS) or statistical potentials,[15,16] Q is adequate only for the fully folded ($Q = 1$) state.[17] For example, for a structured peptide simulated by a transferable force field, some conformations with $Q \approx 0.7$ belong to the denatured state ensemble, and conformations with $Q \approx 0.3$ belong to the folded state.[18,19]

The multidimensionality of the system makes the choice of order parameters for presenting the free-energy surface very important and often leads investigators to compare the results from several different sets. Moreover, the likelihood of hiding essential information concerning the free-energy surface by the commonly used projections has led to a search for alternative methods that are useful for studying protein folding, as well as other complex reactions. One approach that was introduced just 10 years ago is based on disconnectivity graphs[20] (also see ref 21). The unprojected free-energy surface is represented by a disconnectivity graph calculated from an equilibrium folding trajectory with the minimum cut (mincut) or balanced minimum cut (bmincut) procedure.[22] The idea of the method is to group the coordinate sets into free-energy minima, according not to the standard geometric characteristics, but rather to the equilibrium dynamics; that is, the trajectory is used to determine the populations of the states, which provide the relative free energies and the rates of the transition between the states, which yield the free-energy barriers. Application of the method to the β-hairpin of protein G demonstrated that the free-energy surface has multiple low free-energy basins in the denatured state, in

* Corresponding authors. Tel.: +33 390 24 5123. Fax: +33 390 24 5124. E-mail: marci@tammy.harvard.edu (M.K.), caflisch@bioc.uzh.ch (A.C.).
† Laboratoire de Chimie Biophysique.
‡ S.K. and S.M. made equal contributions to this study.
§ University of Zurich.
∥ Harvard University.

10.1021/jp711864r CCC: $40.75
Published on Web 07/01/2008

addition to the native basin, results that complement the analysis of the experimental observation of two-state folding.[23] The same simulation data as were used to reveal the complexity of the denatured state show a relatively smooth free-energy landscape when projected onto a few geometrical coordinates.[22] This result demonstrates that, to obtain a projection that gives an accurate description of the essential aspects of the free-energy surface, different progress coordinates are required.

Projected free-energy surfaces are most useful if they preserve the barriers and minima in the order in which they are encountered during folding/unfolding events. Recently, a new progress coordinate that has some of the desired properties was introduced.[24] It uses the (normalized) partition function of a given region as the progress coordinate and determines the free-energy barriers as a function of the coordinate by a method based on p_{fold}, defined as the probability of reaching the folded state before an unfolded conformation.[25] The result is a one-dimensional projected free-energy profile (FEP) that preserves the barriers between the free-energy basins; given the barriers, the minima can be determined.[24] The method was applied to the β-hairpin of protein G, using root-mean-square deviation (rmsd) clustering, and the conclusions concerning the multiminimum character of the free-energy surface obtained from the disconnectivity graph analysis[22] were confirmed.

It is of interest to apply the methodology just described to a system that is more complex than the β-hairpin. An excellent candidate is the β-sheet miniprotein, called Beta3s.[26] Its structure corresponds to a three-stranded antiparallel β-sheet consisting of two β-hairpins.[27] It has been shown to fold to the native structure determined by NMR spectroscopy[27] in molecular dynamics simulations with a polar hydrogen molecular mechanics potential function modulated by a simple implicit solvent model.[26] Because folding simulations of this system are very fast (for example, close to the melting temperature, the folding time is about 100 ns and requires about 24 h on an Athlon 1.7 GHz computer), many studies have been conducted to elucidate the folding mechanism. Two main folding pathways were observed: one begins with formation of the C-terminal β-hairpin, followed by association of the N-terminal strand on the preformed β-hairpin, and the other follows the symmetry-related pathway (first formation of N-terminal β-hairpin).[26] Conformations in the denatured state of Beta3s were shown to contain a significant amount of non-native contacts,[28] and the folding mechanism of Beta3s had a weak temperature dependence.[29] Moreover, because of the very efficient implicit solvent model, multiple simulations of Beta3s and 32 single-point mutants (totaling 0.65 ms) were performed to directly calculate φ values[30] from folding and unfolding rates extracted from equilibrium folding/unfolding trajectories.[31] More recently, an earlier network analysis of Beta3s[32] was extended to determine free-energy basins.[19] Secondary structure was used to coarse-grain and label the conformations visited in the simulations. The folded state and the most populated free-energy basins in the denatured state were isolated by grouping conformations according to fast relaxation in equilibrium trajectories, a procedure called kinetic grouping analysis (KGA).[19] The comparative application of KGA to Beta3s and a central-strand mutant thereof (W10V) revealed how a single-point mutation can alter the population of native and non-native basins, as well as the relative accessibility of parallel folding pathways.[19] It was shown that only one parameter is required for grouping, namely, the commitment time τ_{commit}, which is chosen as a typical relaxation time within the basins of the investigated system. Any two conformations are grouped into the same basin if they interconvert within the time τ_{commit} with a probability, p_{commit}, of ≥ 0.5. In other words, two conformations are said to be separated by a short kinetic distance if the interconversion between them is fast, which implies that the free-energy barrier between them is low. Kinetic distance is used here, in analogy to the previously introduced term of kinetic closeness,[11] to distinguish what is being described from structural distance.

In the present work, several different, but related, approaches for determining one-dimensional FEPs were applied to Beta3s and the W10V mutant using 20 μs of sampling for each peptide at 330 K, which is slightly above the melting temperature, to obtain adequate sampling of the native and denatured regions of the free-energy surface. The method bpfold, described in ref 24, is based on p_{fold}. As in the balanced mincut method,[22] which finds exact barriers separating individual basins, an extra node is introduced to represent the unfolded state. The extra node is connected to all nodes in the network with a capacity proportional to a Lagrange multiplier λ. For different values of λ, different partitions into two basins with $p_i < 0.5$ and $p_i > 0.5$ are obtained in the bpfold procedure. One approximation to bpfold, referred to as pfoldf (which stands for pfold fast), requires only one value of λ.[24] Two new related approaches were used in this work for comparison. One is called pfoldt and is based on $p_{\mathrm{fold}}(\tau_{\mathrm{commit}})$, which is defined as the probability of reaching the folded state within the time τ_{commit}.[33–35] In the second procedure, the mean first passage time (mfpt) to the native state is used; the procedure is called mfpt. The main difference from pfoldf is that the calculations of the progress variables, $p_{\mathrm{fold}}(\tau_{\mathrm{commit}})$ and mfpt, depend only on the native node; that is, no extra node needs to be added to represent the unfolded state in these procedures. Conversely, only pfoldf is suitable for calculating the barrier between two existing nodes, because only one node can be specified in the mfpt and pfoldt procedures. Further, evaluation of the exact $p_{\mathrm{fold}}(\tau_{\mathrm{commit}})$ values is computationally more expensive than p_{fold} or mfpt calculations, as detailed in the Methods section and the Supporting Information. All three approaches applied here have in common that they encode the kinetic distance to the folded (or any other representative) state and are therefore expected to give similar results. Indeed, the one-dimensional FEPs were found to be very similar and to approximate the exact mincut barrier equally well, underlining the robustness of the methods. The significance of this result is discussed. Further, we compare the results using secondary structure clustering with those obtained with rmsd clustering. It is shown that the barriers from the former tend to be lower than those obtained with the latter; with a proper choice of the rmsd value used for clustering (here found to be 2.5 Å for all-atoms), the resulting Monte Carlo (MC) kinetics agrees with that calculated directly from the trajectories.

Interestingly, a helical state with a statistical weight of about 11% is identified by the three procedures as a free-energy basin separated by a barrier from the rest of the denatured ensemble. The implications of non-native secondary structure content in the denatured state of a β-sheet peptide are briefly discussed. Further, both KGA and one-dimensional FEPs reveal a large and heterogeneous entropic region (weight of 33%) that is separated by a barrier of less than k_BT from the native state (weight of 35%). The single-exponential behavior of Beta3s folding is shown to be due to the similar free-energy barriers to exit from the non-native enthalpic traps (total population of about 20%) or from the helical basin (weight of 11%), which is primarily stabilized by its entropy.

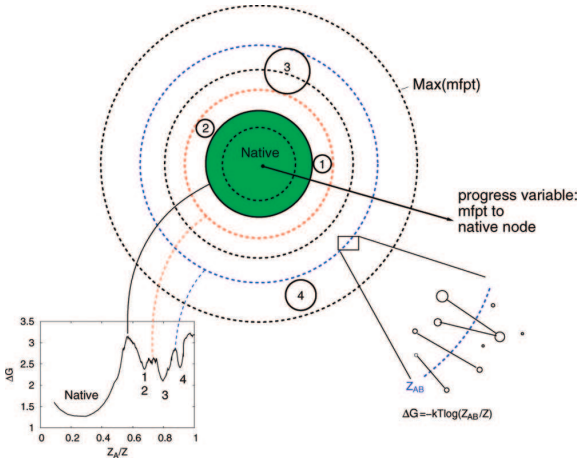


Figure 1. Schematic illustration of the one-dimensional FEP procedure using mfpt as the progress variable. Each of the four solid circles represents a free-energy basin, and the concentric dashed circles represent values of mfpt. For each value of mfpt_c between 0 (native node) and max(mfpt), a point in the profile is obtained. ΔG of the fraction of links crossing the cutting surface at mfpt = mfpt_c (bottom right) is plotted as a function of the relative partition function Z_A/Z (bottom left), where the set A contains all nodes with mfpt<mfpt_c. Basins 1 and 2 overlap because they have the same mfpt distance from the native state and are therefore not separated in the unfolded part of the profile. Note that the same illustration is valid if mfpt is replaced by p_{fold} or $p_{\text{fold}}(\tau_{\text{commit}})$, with the only difference being that the distance from native would decrease from 1 (native) to 0.

Methods

Molecular Dynamics Simulations. All simulations and most of the analysis of the trajectories were performed with the program CHARMM;[36] the rest of the analysis was done with the program WORDOM,[37] which is particularly efficient in handling large sets of trajectories. The designed 20-residue peptide Beta3s[27] (Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17}-Ile_{18}-Tyr_{19}-Thr_{20}) and its W10V mutant were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[38] with a default cutoff of 7.5 Å for the nonbonding interactions). A mean-field approximation based on the solvent-accessible surface (SAS) was used to describe the main effects of the aqueous solvent.[39] It has been shown previously that this model for the solvated Beta3s peptide yields reversible folding at 330 K to the NMR conformation, irrespective of the starting structure; 23 of the 26 nuclear Overhauser effect (NOE) constraints are satisfied.[26] Moreover, despite the neglect of collisions with water molecules (frictional effects) in the simulations with the implicit solvent model, the *relative* rates of folding for different secondary structural elements are comparable to the values observed experimentally; i.e., helices fold in about 1 ns,[40] β-hairpins in about 10 ns,[40] and triple-stranded β-sheets in about 100 ns,[31] compared to experimental values of ~0.1,[41] ~1,[41] and ~10 μs,[27] respectively. For Beta3s and the W10V mutant, 10 molecular dynamics runs of 2 μs each with different initial distributions of velocities were performed with the Berendsen thermostat (coupling constant of 5 ps) at 330 K, which is slightly above the melting temperature of Beta3s.[29] A time step of 2 fs was used, and the coordinates were saved every 20 ps, for a total of 10^6 snapshots for each system. This required three weeks on a 20-CPU cluster. Using explicit water simulations, it would have been much more time-consuming to obtain the 40 μs of simulation time required to sample a statistically significant number of equilibrium folding/unfolding transitions.

Coarse-Graining and Equilibrium Kinetic Network (EKN). For the purpose of using the finite-time simulation data to obtain free-energy surfaces, it is necessary to coarse-grain the snapshots in some way because each conformation is visited only once; the trajectory, per se, is nothing but a long string of configurations. There are several meaningful methods for clustering individual coordinate sets from the trajectory to obtain coarse-grained conformations, and different approaches are likely to be most useful for different types of analysis. For a system such as Beta3s or a β-hairpin, rmsd and secondary structural coarse-graining are obvious possibilities.[22,32,42] The coarse-graining used in this work is based on secondary structure strings.[43] A "coarse-grained conformation" or node is a single string of secondary structure; for example, the most populated conformation of Beta3s, which corresponds to the native state, is −EEEES-SEEEEEESSEEEE−.[32] There are eight possible "letters" in the secondary structure "alphabet": H, G, I, E, B, T, S, and −, standing for α-helix, 3_{10}-helix, π-helix, extended, isolated β-bridge, hydrogen-bonded turn, bend, and unstructured, respectively.[43] Because the N- and C-terminal residues are always assigned as unstructured,[43] a 20-residue peptide can, in principle, assume $8^{18} \approx 10^{16}$ conformations. We note that there is no relation between the Hamming distance (number of different entries, i.e., letters, in two strings of equal length) and the kinetic distance (see also the Results section). The secondary-structure-based coarse-graining is used to permit comparison with earlier work.[19] It has the advantage over approaches based on the rmsd of the atomic coordinates of being more efficient because it scales with the number of snapshots whereas rmsd is a pairwise measure. Also, each node is uniquely defined by its secondary structure string, which serves as a useful conformational "label". However, as we show here, secondary-structure-based coarse-graining can, in some cases, lead to overlapping of regions that are distant in terms of rmsd (i.e., far from each other in configuration space), which can result in barriers for the basins that are too low; this is the case for Beta3s and is referred to as "pseudotunneling" hereafter.

The number of snapshots with a given secondary structure string i is called the weight of the node and is denoted as $\tilde{w}$. The statistical weight w of a node is given by $w = \tilde{w}/N$, where $N = 10^6$ is the total number of snapshots. In the same way, the links, which are direct transitions sampled along the MD trajectory, are weighted by n_{ji}, defined as the number of times a snapshot in node i is followed by a snapshot in node j. As mentioned above, snapshots were saved every 20 ps, which is therefore the time interval of a direct transition. The resulting equilibrium kinetic network (EKN)[24,44] is an undirected, weighted graph where the edge capacity from node j to node i in the network, c_{ij}, is proportional to the number of direct transitions from j to i at equilibrium. Detailed balance can be "imposed", i.e., $c_{ij} = c_{ji} = (n_{ij} + n_{ji})/2$. The transition probabilities can then be calculated as $p_{ij} = c_{ij}/\Sigma_k c_{kj}$.

For a node i in the EKN, the partition function is $Z_i = \Sigma_j c_{ij}$. If the nodes of the network are partitioned into two groups A and B, then $Z_A = \Sigma_{i\in A} Z_i$, $Z_B = \Sigma_{i\in B} Z_i$, $Z_{AB} = \Sigma_{i\in A, j\in B} c_{ij}$, and the free energy of the barrier between the two groups is $-kT \log(Z_{AB}/Z)$, where Z is the partition function of the full network (Figure 1).

One-Dimensional FEP. The pfoldf procedure to determine the one-dimensional FEP was published previously;[24] two additional procedures, pfoldt and mfpt, are introduced here.

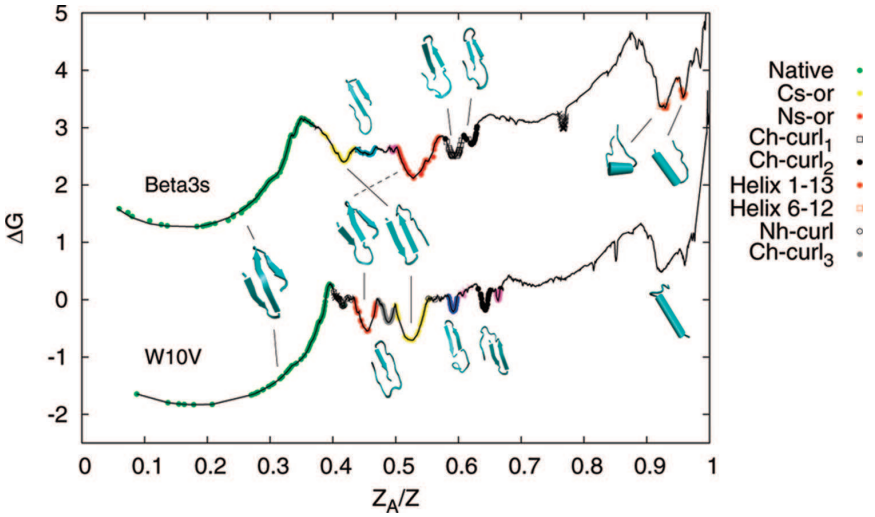


Figure 2. pfoldf-calculated FEP of Beta3s and its single-point mutant W10V, whose plot is shifted by −3 kcal/mol to avoid overlap of the curves. The progress coordinate is the relative partition function for different values of p_c (see text and Table 1). Symbols represent the secondary structure strings with ≥100 snapshots; they are colored according to the basins identified by the KGA.[19] Conformations in the most populated basins are shown as ribbon diagrams, where β-strands are represented by arrows and helices by cylinders. Abbreviations: Ns-or, N-terminal strand out of register and folded C-terminal hairpin; Cs-or, C-terminal strand out of register and folded N-terminal hairpin; Nh-curl, curl-like conformation with folded N-terminal hairpin; Ch-curl, curl-like conformation with folded C-terminal hairpin. Note that Ns-or conformations have a topology similar to that of the folded structure but non-native orientations of the side chains in the N-terminal strand. The Z_A/Z coordinate value of the first data point (leftmost green circle) is the relative weight of the most populated secondary structure node.

TABLE 1: Overview of the Four Procedures Discussed in the Text for Determining FEPs[a]

	bpfold[b]	pfoldf[c]	pfoldt[d]	mfpt[e]
for barrier to exit from a basin	yes	yes	yes	yes
for barrier between two basins	yes	yes	no	no
number of target nodes	2	2	1	1
extra node	yes	yes	no	no
system of equations	$p_i = \Sigma_j p_{ji} p_j$	$p_i = \Sigma_j p_{ji} p_j$	(see *Supporting Information*)	$\text{mfpt}_i = \Delta t + \Sigma(p_{ji} \times \text{mfpt}_j)$
boundary condition(s)	$p_A = 1$, $p_B = 0$	$p_A = 1$, $p_B = 0$	$p_A = 1$	$\text{mfpt}_A = 0$
progress variable	p_{fold}	p_{fold}	$p_{\text{fold}}(\tau_{\text{commit}})$	mfpt
Z_A/Z evaluation	$p_{\text{fold}} > 0.5$	$p_{\text{fold}} > p_c$; $0 \le p_c \le 1$	$p_{\text{fold}} > p_c$; $0 \le p_c \le 1$	$\text{mfpt} < \text{mfpt}_c$; $0 \le \text{mfpt}_c < +\infty$
number of λ values	many	1	0	0
ref	24	24	this work	this work

[a] Three approaches related to bpfold (i.e., pfoldf, pfoldt, and mfpt) were used in this work. The pfoldf and mfpt procedures are available in the MD analysis package WORDOM.[37] [b] bpfold: The balanced pfold requires an extra node if no representative node of the unfolded state exists. This approach yields the most accurate description of the barrier to exit a basin. [c] pfoldf: The pfold fast evaluation is an approximation of bpfold that uses only one value of the Lagrange multiplier λ. [d] pfoldt: The p_{fold} between two nodes is replaced by $p_{\text{fold}}(\tau_{\text{commit}})$, which is the probability of reaching a target node within the commitment time τ_{commit}. The extra node and Lagrange multiplier are not required. pfoldt is suitable for the calculation of unfolding profiles, but cannot be employed for the detection of a barrier between two basins. [e] mfpt: Similarly to pfoldt, the mean first passage time to a target node does not require an extra node, and therefore, it applies to the same cases as pfoldt.

Table 1 lists details of all of the procedures used to calculate FEPs in this study.

pfoldf. Given the EKN and two nodes A and B, the p_{fold} value of node i, p_i, is found as the solution of the equation $p_i = \Sigma_j p_{ji} p_j$ with boundary conditions $p_A = 1$ and $p_B = 0$ (A is considered to be the "native node" and B the "denatured node"). The system of equations can be solved efficiently numerically by iterative multiplication of the vector p_j by the matrix p_{ji}.[24]

To determine the FEP relative to a chosen node A, node B is considered to be the representative node of everything not belonging to the basin of A. However, in many systems, a node such as B does not exist, because there are multiple basins and/or an entropic state that cannot be represented by a single node. Both occur in the peptides investigated in this study. Thus, as in the balanced minimum-cut procedure,[22] an extra node B is introduced and connected to all nodes in the network with capacity $\lambda\tilde{w}$, where λ is a Lagrange multiplier. The p_{fold} calculations are performed on the EKN with the extra node, and in the pfoldf procedure, the nodes are sorted according to their p_{fold} values using only one value of λ. The assumption in this procedure is that the order of the nodes does not change for different values of λ.[24] Each value p_c between 0 and 1 can then be used to cut the network into set A containing all nodes with $p_{\text{fold}} > p_c$ and set B containing the nodes with $p_{\text{fold}} < p_c$. For each cut, a point $[x = Z_A/Z, y = -kT \ln(Z_{AB}/Z)]$ of the FEP is obtained; Z_A/Z is used as the progress coordinate, and Z_{AB} is the number of EKN transitions between the two sets. Note that p_{fold} is the progress variable used to divide the configuration space, with pfoldf evaluation based on this variable. Moreover, this procedure and those described below do not require any special treatment of low-population nodes (i.e., secondary structure strings with only one or a few snapshots) because they are automatically grouped on the same side of the barrier as the reference node A if they satisfy the condition $p_{\text{fold}} > p_c$.

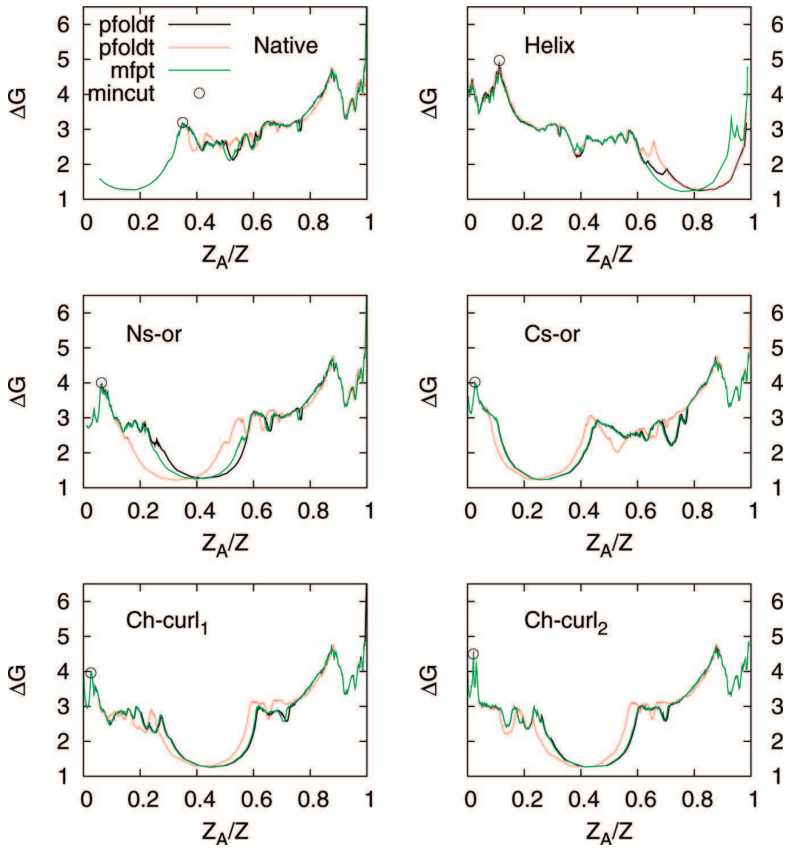


Figure 3. One-dimensional FEPs calculated using the kinetic distance from individual basins of Beta3s. Notably, the barriers separating the reference state from the rest are almost identical for pfoldf with $\lambda = 0.0001$ (black), pfoldt (red), and mfpt (green), and all three procedures yield good approximations to the exact barrier height calculated with mincut[22] (open circle). pfoldt was calculated with different τ_{commit} values, namely, 20 ns (native), 40 ns (Ns-or), 100 ns (Cs-or, Ch-curl$_1$, Ch-curl$_2$), and 200 ns (helical). Except for the shift of some of the minima after the first barrier, the three FEPs are very similar, which indicates that the approach is robust upon the choice of the progress variable encoding for kinetic distance from a reference state.

pfoldt. The extra node required by the pfoldf procedure is not necessary if p_{fold} is calculated not between two representative nodes A and B, but rather with a commitment time τ_{commit}, referred to as $p_{fold}(\tau_{commit})$ and defined as the probability of reaching A within τ_{commit}.[33−35] The calculation of $p_{fold}(\tau_{commit})$ values for all nodes in the EKN (with the initial boundary condition $p_A = 1$) is more complex than that for pfoldf; details are given in the Supporting Information. Once $p_{fold}(\tau_{commit})$ has been evaluated for all nodes, the procedure is the same as for pfoldf: The nodes are sorted according to $p_{fold}(\tau_{commit})$ values and split into sets with $p_{fold}(\tau_{commit}) > p_c$ and $p_{fold}(\tau_{commit}) < p_c$. For each p_c between 0 and 1, the pair $[Z_A/Z, -kT \ln(Z_{AB}/Z)]$ is a point on the FEP. The choice of τ_{commit} has to be long enough to assign nonzero $p_{fold}(\tau_{commit})$ values to nodes that are kinetically very far from the node under consideration (i.e., the native node or any other node of interest) to resolve all other states. For this purpose, very long commitment times are appropriate. However, it is computationally most convenient to choose τ_{commit} values as short as possible. Thus, for each basin, one starts with a short τ_{commit} and increases it until the whole profile is covered, as illustrated in Figure S1 of the Supporting Information. For instance, 20 ns is long enough if $p_{fold}(\tau_{commit})$ is calculated with respect to the native node, but $\tau_{commit} = 200$ ns is needed if the values are calculated with respect to a node in the helical region. All pfoldt profiles shown in Figure 3 below were produced with values in this range (20−200 ns). Typically, the upper limit of τ_{commit} is on the order of the overall relaxation time ($\tau_{folding} + \tau_{unfolding}$) of the system, which is about 200 ns for Beta3s.

mfpt. Another variable used in this work to project the free energy is the mean first passage time (mfpt; see Figure 1) to node A (representative nodes of significantly populated basins were used in Figure 3, but any node can be used as a reference).[45] Given the original EKN (i.e., without an extra node), the mfpt of node i is the solution of the equation $\text{mfpt}_i = \Delta t + \Sigma(p_{ji} \times \text{mfpt}_j)$ with initial boundary condition $\text{mfpt}_A = 0$.[46] The time step Δt corresponds to the saving frequency of 20 ps; that is, the mfpt of a node is defined as one time step plus the weighted average of the mfpt values of its adjacent nodes. In contrast to the other progress variables, the mfpt has an explicit time dependence through the occurrence of the time step in the equations. The resulting system of linear equations differs from that of pfoldf by the Δt constant and the boundary conditions; in pfoldf, the boundary conditions are $p_A = 1$ and $p_B = 0$, whereas there is only one condition $p_A = 1$ or $\text{mfpt}_A = 0$ for pfoldt or mfpt, respectively. Therefore, both pfoldf and mfpt equations can be solved with the same efficiency by iterative multiplication. Similarly to pfoldt, mfpt does not require an extra node, because mfpt is defined not between a pair of nodes, but only with respect to one selected node. To calculate the FEP, the nodes are sorted according to their mfpt values. For any mfpt_c between 0 and max(mfpt), a point $[Z_A/Z, -kT \ln(Z_{AB}/Z)]$ on the FEP can be calculated, where A is the set of

all nodes with $mfpt_i < mfpt_c$ and B is the set of nodes with $mfpt_i > mfpt_c$ (Figure 1).

Implementation. In practice, the procedure to calculate the one-dimensional FEP consists of four steps: (1) Detailed balance is imposed on the equilibrium kinetic network (EKN), i.e., $c_{ij} = c_{ji} = (n_{ij}+n_{ji})/2$, where n_{ij} is the number of direct transitions (i.e., transitions between two MD snapshots separated by the time interval Δt, which is the inverse of the MD saving frequency) from node j to node i. The transition probabilities are then calculated as $p_{ij} = c_{ij}/\Sigma_k c_{kj}$. (2) The system of equations with appropriate boundary condition(s) is solved numerically. (3) Nodes are sorted according to increasing values of mfpt or decreasing values of pfoldf or pfoldt; for each value of the progress variable, the relative partition function Z_A and the cut Z_{AB} are calculated. (4) The individual points on the profile are evaluated as $[x = Z_A/Z, y = -kT \ln(Z_{AB}/Z)]$.

Identification of Basins. The kinetic grouping analysis (KGA) groups conformations according to fast relaxation at equilibrium.[19] More explicitly, two coarse-grained conformations are grouped if, along the molecular dynamics trajectory, their snapshots interconvert in more than 50% of the cases within a commitment time τ_{commit}, which represents a typical relaxation time within basins of the investigated system; the value used in this study was 1 ns. The basins obtained by KGA can be compared to those isolated from FEPs. To isolate a basin with a FEP, the unfolding profile from a node in that basin (usually its most visited node) is plotted, as shown in Figure 3. In practice, the procedure is the same as that used with the native basin as the reference, except that the native node is replaced by the new node. All nodes lying on the left of the cut at the first barrier correspond to the basin. Basins lying on the right of the first barrier are potentially overlapping (Figure 1), so each basin requires a separate unfolding profile.

Transition Disconnectivity Graph (TRDG). The TRDG is a variant of the free-energy disconnectivity graph, which provides an unprojected representation of the free-energy surface.[44] The partition function of the free-energy barrier separating states i and j, Z_{ij}, is equal to the value of the mincut between the states in the network, which can be calculated by the Ford–Fulkerson algorithm.[47] After the mincuts (i.e., the free-energy barriers) between every pair of nodes have been calculated, which can be done with only $n - 1$ total mincuts for n nodes by using the Gomory–Hu algorithm,[48] the TRDG is constructed to obtain a detailed representation of the free-energy surface. Following Becker and Karplus[20] and using the relation $F_{ij} = -kT \ln(Z_{ij})$, one starts with the largest Z_{ij} value (smallest F_{ij} value) and successively connects states in order of decreasing Z_{ij} (increasing F_{ij}). The TRDG is useful for visualizing basins containing representative nodes (enthalpic basins), but basins that have no such nodes (entropic basins) are not visible. To resolve such basins, one has to use either the balanced mincut procedure[22] or the free-energy profiles discussed above.

Results

All analyses are based on a set of of 10 2.5-μs equilibrium simulations at 330 K started from the folded state. The first 0.5 μs of each run was neglected so that a total simulation time of 20 μs was sampled for each of the two peptides (see Methods). The wild-type Beta3s peptide visited 262 433 conformations (unique strings of secondary structure) with a total of 534 383 direct transitions. The W10V mutant visited 245 032 conformations with a total of 476 721 direct transitions. In Beta3s (W10V), only 62 446 (56 118) conformations were visited more than once. The most populated conformation of Beta3s and its W10V mutant had statistical weights of 5.6% and 8.8%, respectively. It was the three-stranded antiparallel β-sheet with type II′ turns at residues 6–7 and 14–15 (secondary structure string –EEEESSEEEEEESSEEEE–), which corresponds to the native state determined by NMR spectroscopy.[27] Totals of 120 and 105 folding events (i.e., visits to the native node) were observed for Beta3s and W10V, respectively, with an average folding time of about 0.1 μs for both peptides.

FEP of Beta3s and W10V. The pfoldf FEP, based on p_{fold} values, was calculated by projecting the free energy on the relative partition function Z_A/Z, which is a progress coordinate that increases monotonically with the distance from the reference state;[24] see Methods. pfoldf finds approximate barriers between the reference state and the denatured state, whereas exact values can be obtained by the mincut procedure.[22] An essential attribute of Z_A/Z is that it takes into account all routes from the initial state to the final state without any prejudgment as to the geometric coordinates or pathways involved, so that the FEP is determined by an unbiased procedure. Figure 2 shows the results for both Beta3s and W10V. Three main regions are identified on the FEP of Beta3s when the most populated (native) node is used as reference: native state ($Z_A/Z < 0.35$), denatured state with several enthalpic subbasins ($0.35 \leq Z_A/Z \leq 0.88$), and helical basin ($Z_A/Z > 0.88$). The pfoldf FEP procedure yields an accurate description of the reference basin (native state in Figure 2), whereas there can be some overlap between different basins after the first barrier (i.e., for $Z_A/Z > 0.35$). Overlap occurs whenever nodes belonging to different basins have similar p_{fold} values (Figure 1). Unfolding profiles (see Methods) are able to fully resolve all the basins. Such unfolding profiles from representative nodes in the subbasins of Figure 2 are plotted in Figure 3 to accurately characterize each basin by eliminating overlap with other regions. This makes possible the determination of accurate barriers between the respective basins and the exact population of each basin. In each unfolding profile of Figure 3, the exact value of the barrier, calculated by mincut,[22] is indicated by an open circle. The plots show that pfoldf approximates the barriers very well for the Beta3s system. (For a description of the pfoldt and mfpt results, see below.) Notably, the most populated basins isolated by this unbiased procedure[24] correspond to those in Table 1 of ref 19 and are quantitatively compared in Table 2 of the present work. Figure 4 qualitatively illustrates the nodes, the basins, and their connectivity in the conformational state network.[32] The basins determined by pfoldf are discriminated by different colors and shapes in the network, where brown nodes belong to the entropic state. Note that the colors in the network (i.e., basins defined by pfoldf) are in good agreement with the network in Figure 1 of ref 19 for KGA results. Table 3 contains effective and free-energy values of the six basins and the entropic state. Clearly, there are low-enthalpy, low-entropy basins (native, Ns-or, Cs-or, and the two curls), as well as high-enthalpy, high-entropy basins (helical and entropic states); the origin of the high entropy of the helical basin is discussed below. Except for the helical basin, the agreement between the FEP procedures and the KGA[19] is very good, and thus, the two approaches validate each other. Furthermore, essentially identical basins are isolated by either of the two procedures using secondary structure or all-atom 2.5-Å rmsd coarse-graining (as shown in Table S-I of the Supporting Information for pfoldf).

Although the Beta3s and W10V profiles are very similar, there are subtle differences between them. Importantly, the native state is less stable in Beta3s than in W10V (35% vs 39.5%), the helical basin is slightly more populated in Beta3s than in W10V

TABLE 2: Results of the KGA and pfoldf Procedure for the Most Populated Basins of Beta3s[a]

basin		weight		
heaviest node	name/color	KGA (%)	pfoldf (%)	similarity[b] (%)
−EEEESSEEEEEESSEEEE−	native	36.4	35.0	99.8
−EEEESTTEEEEESSEEEE−	Ns-or	7.4	6.2	99.99
−EEEESSEEEEESSSEEEE−	Cs-or	3.6	2.6	99.8
−−−SSGGG−−−EESSEETT−	Ch-curl$_1$	3.6	2.8	99.5
−−−SSGGG−EESSTTTTEE−	Ch-curl$_2$	2.4	2.1	98.8
−HHHHHHHHHHHHS−−−−−−	helical[c]	2.1 (11.6)	11.2	99.4 (96.4)
−−EESSSEEEEEESSEEEE−	cyan	1.9	1.5	97.9
−−−SSGGG−EEESSSEEEE−	white rectangle	1.2	1.1	98.8
−−−SSSS−−EEESTT−EEE−	white diamond	0.9	0.9	96.3
−−−−STT−−−EEESSEEEE−	magenta	0.8	0.5	99.7
−	entropic-1[d]	33.8	32.7	76.0
−	entropic-2[e]	27.2	32.7	93.0

[a] The pfoldf, pfoldt, and mfpt procedures determine the basin populations by plotting the FEP from the representative (most populated) node for each basin as shown in Figure 3. The cut at the first barrier defines each basin. Note that the pfoldf procedure yields essentially identical basins using clusters obtained by secondary structure or all-atom 2.5-Å rmsd coarse-graining (see Table S-I in the Supporting Information). [b] The similarity value is calculated as the intersection of two corresponding basins (i.e., that obtained using KGA and that using pfoldf), normalized to the one with the lower population. Note that pfoldf, pfoldt and mfpt populate the investigated basins equally (similarity at least 99%, not shown). [c] Using KGA, the population of the largest helical basin is only 2.1%, while the ensemble of strings with 4, 5 or 6 consecutive G, H and I's, respectively, populates a total of 11.6%, which is very close to the helical basin population obtained by pfoldf. [d] The entropic-1 state is defined for KGA as conformations not belonging to basins with more than 1% population, and for the pfoldf procedure as conformations not belonging to one of the significant basins of the profile. [e] The KGA entropic-2 basin does not include the helical conformations.

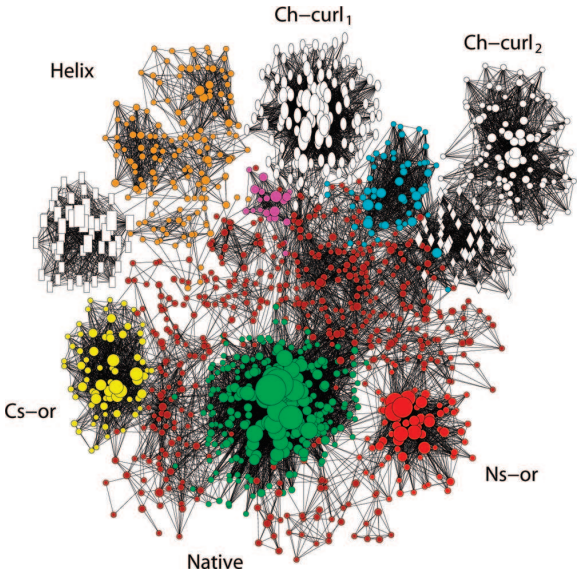


Figure 4. Conformational space network of Beta3s. Each node (i.e., conformation) of the network represents a secondary structure string. The surface of each node is proportional to its statistical weight, and only the 1430 nodes containing at least 40 snapshots are shown to avoid overcrowding.[19] Nodes are colored according to the basins isolated with pfoldf FEPs (Figure 2 and Table 2). White nodes are populated significantly only in Beta3s and not in the single-point mutant W10V;[19] the different white symbols are assigned according to pfoldf (Table 2). Note that the helical basin is assigned a single color when isolated according to the pfoldf FEP and multiple colors by KGA,[19] as explained in the text. Brown is the color of some of the nodes in the entropic basin of which most nodes are not shown because they contain less than 40 snapshots each.

(11.2% vs 8.8%), and there is a difference in the relative stability of nonhelical misfolded species such as Ns-or (N-terminal strand out of register and folded C-terminal hairpin, basin populations of 6.2% vs 3.6%) and Cs-or (C-terminal strand out of register and folded N-terminal hairpin, basin populations of 2.6% vs 4.9%). These statistical weights were calculated by defining the basins as described in the Methods section. Note that corresponding basins in the two systems can occur at different Z_A/Z positions in the profiles of Figure 2. Also, the FEPs reveal basins that are visited in only one of the two peptides. Such basins are illustrated in white in Figure 4.

Figure 5 shows the negative logarithm of the probability of the first passage time (fpt) to the native node (i.e., the folding time). The plot has a minimum at about 100 ns, which corresponds to the folding time. The fast folding values correspond to configurations that start in the native state (fpt < 1 ns), and the slower folding ones are configurations that start in the denatured state (fpt > 1 ns). The fpt plot shows a very simple behavior, such as is expected for a two-state system, despite the multiminimum character of the free-energy surface.

FEPs Calculated with pfoldt and mfpt. The pfoldf analysis of systems without a representative node in the denatured state is based on the introduction of an extra node[24] that is linked with a small capacity λ (typically 0.01 or lower) to all nodes in the network. Here, p_{fold} with a commitment time of $p_{fold}(\tau_{commit})$ and the mean first passage time (mfpt) to the reference node as progress variables are introduced to plot the FEPs. The respective procedures, called pfoldt and mfpt, have the advantage that no additional node needs to be introduced.

The pfoldt- and mfpt-calculated FEPs are in good agreement with those obtained by the pfoldf procedure (Figure 3). The pfoldf, pfoldt, and mfpt barriers separating significantly populated basins are almost identical for all procedures and both peptides. Further, pfoldt and mfpt, as well as pfoldf (see above), yield good approximations to the exact barriers (open circles in Figure 3). The six basins shown in Figure 3 share more than 99% of their conformations when isolated by the three methods (not shown), indicating that the results are robust. This suggests that the choice between pfoldf, pfoldt, and mfpt can be made according to convenience. The cut between two well-defined regions with representative nodes has to be calculated by pfoldf, because it is the only procedure of the three discussed here in which two input nodes are used, but pfoldt and mfpt are the more straightforward choices if only one representative node exists, such as when an unfolding profile is calculated. Solving

TABLE 3: Energetic and Entropic Contributions to Basin Stability

pfoldf basin	weight (%)	$\langle E\rangle^a$ (kcal/mol)	$\langle \Delta E\rangle^b$ (kcal/mol)	ΔF^c (kcal/mol)	$-T\Delta S = \Delta F - \langle \Delta E\rangle$ (kcal/mol)	barrier[d] secondary str (kcal/mol)	barrier[d] rmsd (kcal/mol)
native	35.0	−7.2	0	0	0	1.9	2.5
Ns-or	6.2	−6.4	0.8	1.1	0.3	1.2	1.7
Cs-or	2.6	−3.3	3.9	1.7	−2.2	0.8	1.6
Ch-curl$_1$	2.8	−10.1	−2.9	1.7	4.6	1.0	3.2
Ch-curl$_2$	2.1	−10.1	−2.9	1.9	4.8	1.2	2.6
helical[e]	11.2	2.2	9.4	0.8	−8.6	1.4	1.3
entropic[e]	32.7	2.9	10.1	0.0	−10.1		

[a] Average effective energy, that is, the sum of CHARMM param19 force field[38] and the SAS solvation model[39] contributions. The average was calculated over all snapshots in the clusters (i.e., secondary structure strings) belonging to the basin determined by the pfoldf procedure. The error of the average effective energy is less than 0.5 kcal/mol as estimated by the difference of the mean values calculated on one-half of the sample, i.e., two segments of 10 μs each. [b] Average effective energy relative to the folded state. Note that, in any force field, the absolute value of the effective energy E is arbitrary, and only ΔE values relative to a reference state are meaningful. [c] Free energy relative to the folded state calculated as $\Delta F = -kT \ln(\text{weight}/35.0)$, where 35% is the weight of the native basin as isolated by pfoldf. [d] Barrier to exit individual pfoldf basin, calculated for the secondary structure and 2.5-Å rmsd coarse-graining. [e] These basins are stabilized mainly by entropy, as indicated in the table.

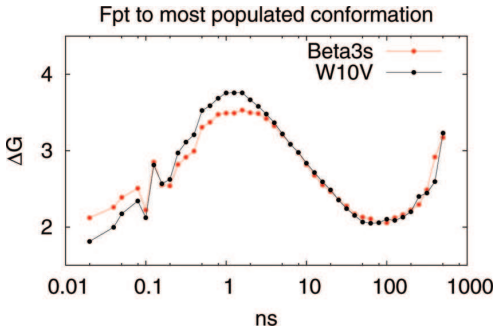


Figure 5. Profile obtained using fpt as the progress variable and calculated as $\Delta G = -k_B T \ln[P(\text{fpt})]$ on the bins, whose size increases exponentially (10 bins/decade) for better resolution of the different timescales.

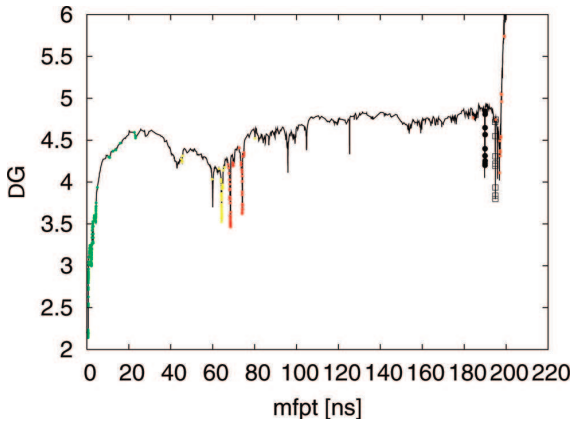


Figure 6. Beta3s unfolding FEP calculated for the directed network (see Methods) using mfpt as a progress coordinate *and* a progress variable, which can be obtained by the transformation $x \rightarrow \text{mfpt}[x(Z_A/Z)]$. As in Figure 3, individual basins are colored according to basins isolated from pfoldf FEPs. Interestingly, the similar mfpt values for the enthalpic traps and the spread of only about 3 between mfpt values of enthalpic traps and the helical basin are consistent with the single-exponential behavior of folding (see the Results section). The 2.5-Å rmsd clustering is used here because the corresponding FEP calculated for the secondary structure coarse-graining is affected by pseudotunneling (see text and Figure S2 in the Supporting Information).

the system of equations for the pfoldf and mfpt procedures is of the same complexity, because the only differences are the boundary conditions and the use of the time constant Δt that is added in the mfpt equations (see Methods). On the other hand, the pfoldt method requires some precalculations before the iterative solution of the equation can be performed and is therefore more complex (see Supporting Information). An application of the mfpt analysis is that it can be used as the progress coordinate (instead of Z_A/Z) to obtain a FEP with all basins separated from the native one by a distance in time units. The relation between the profiles projected on Z_A/Z and on mfpt is a nonlinear transformation $x \rightarrow \text{mfpt}[x(Z_A/Z)]$, where $x(Z_A/Z)$ assigns nodes to each position on Z_A/Z; that is, the mfpt that was originally used to rank on the Z_A/Z axis is now directly assigned to the nodes. Figure 6 clearly shows mfpt values of individual basins. Interestingly, it also illustrates the origin of the single-exponential behavior (see Single-Exponential Kinetics of Folding), which is a spread of only a factor of 3 in the mfpt values (from about 60−80 ns for Ns-or and Cs-or to about 190−200 ns for Ch-curl$_1$, Ch-curl$_2$, and helical). However, as can be seen by comparing Figure 3 with Figure 6, there is more overlap of non-native basins using mfpt rather than Z_A/Z as the progress coordinate.

Helical Basin. A previous study[32] suggested that the denatured state ensemble of Beta3s is highly heterogeneous and includes enthalpic traps as well as conformations with partial helical structure; the latter form the helical basin. Notably, in the FEPs of Figures 2 and 3, the entire "helical" region is identified and shown to be separated by a high barrier (at Z_A/Z of about 0.88) from the rest of the denatured state, which extends from 0.35 to 0.88 (see above). The helical region shows the main difference with respect to the KGA results[19] and is indicated in Figure 4. KGA correctly identified the two most populated free-energy subbasins (−HHHHHHHHHH-HHS−−−−−− and −−TT−−HHHHHHHSS−−−−− with populations of 1.9% and 1.6%, respectively) within the helical state of Beta3s (Figure 2). The commitment time of 1 ns used in the previous work[19] was too short to group all helical structures into one basin, because the helical basin is divided into various subbasins separated by barriers, as can be seen in the helical unfolding profile of Figure 3. These barriers prevent the system from rapid equilibration between all helical structures. A larger commitment time of 5 ns, however, is able to identify the entire helical basin (Supporting Information of ref 19). These results show that the definition of a basin involves the choice of "resolution". Both the commitment time of KGA and the height of the barrier in the FEP analysis, above which

one considers a basin as separated, correspond to the "lens" with which the free-energy surface is analyzed. For each choice of a minimum barrier in the FEP procedures, there exists a commitment time for the determination of the corresponding basins with KGA. However, defining a minimum barrier height is more transparent than choosing a commitment time, which is initially extracted from the fpt plot of Figure 5 and then varied to obtain the desired resolution. In either case, there is some arbitrariness in defining a basin per se.

The helical state is the right-most basin in all five unfolding profiles from the nonhelical basins in Figure 3 (see also Figure S2 in the Supporting Information). This observation is consistent with the high barrier that has to be overcome to enter the helical basin from the rest of the conformational space, as the barriers generally appear in increasing order along the Z_A/Z progress coordinate.

Interestingly, as shown in the thermodynamic analysis presented in Table 3, the helical basin has a high energy and is entropically stabilized. This is because the strings associated with the helical basin have unstructured residues that do not make hydrogen bonds. In the helical basin, 78.4% and 7.7% of the snapshots have more than 5 and 10 unstructured residues, respectively; that is, they belong to strings with more than 5 and 10 "−" letters. The corresponding percentage values for the entropic region are 79.7% and 15.2%, respectively. As a basis of comparison, the native state has only 2.9% and 0.003% of its snapshots in strings with more than 5 and 10 unstructured residues, respectively. Moreover, the numbers of different secondary structure strings in the helical, entropic, and native states are 57 134, 193 666, and 2672, respectively.

Simplified Network (SEKN) and the Role of the Entropic State. In previous analyses of folding simulations, it has been found useful[22,24] to construct a highly simplified network that shows only the main basins and their connectivity as a complement of the detailed conformational space network (Figure 3). Figure 7 shows such a network with the free-energy basins isolated by pfoldf. It includes the heterogeneous entropic state (dashed surface in Figure 7), which is made up of all conformations not belonging to any of the 10 basins that appear as a part of the network; the latter all have barriers higher than 0.3 kcal and significant partition function (≥0.5%). The so-called entropic state, by contrast, is composed mainly of nodes that are visited only once, or at most a few times, so that it is not a "true" basin in the sense used for the other basins. Overall, at the chosen simulation temperature (330 K), the denatured state of Beta3s consists of the entropic state (populated at 33%), a helical basin (populated at 11%) and eight metastable enthalpic traps (populated at 0.5−6%). Although the free-energy basins were selected with the pfoldf procedure on the full EKN, the links in Figure 7 show the number of transitions, sampled along the molecular dynamics trajectory, between the most populated secondary structure string (bottom) of a given pfoldf basin i and the bottom of another basin j (or the same basin) through the entropic state only (i.e., without visiting another enthalpic basin). Each of the pfoldf profiles in Figure 7 was calculated using only the basin under consideration, the entropic state, and the native state while neglecting the other basins. In this way, barriers involved in the transitions between individual enthalpic basins and the native state are described accurately; see also the Supporting Information, section D. Note that the left-most profile in Figure 7 (i.e., pfoldf-calculated FEP from the entropic state) considers only the entropic and native states, which account for almost 70% of the total weight. All profiles show a barrier of only about 0.5 kcal/mol from the entropic state toward the native state, which is near the limit of what has been termed barrierless or downhill folding.[49–54]

The complexity of the SEKN, particularly for the denatured state, suggests that a detailed analysis would be useful to obtain a more complete understanding of the folding behavior. Figure 7 shows the number of "direct" transitions (i.e., without visiting the native state) between identified basins, which can be compared with the number of transitions from each of the basins to the native state. In most cases (the pair cyan and Ns-or is an exception), the direct transitions are rather rare compared to the number of transitions connecting each basin to the native state. However, the total number of transitions connecting the non-native basins to other non-native basins without passing through the native state is of the same order of magnitude as the number of transitions connecting each of the non-native basins to the native state.

If the equilibrium trajectories are followed, they make clear that direct transitions overwhelmingly go through the entropic basin from one of the defined basins to another. This is in accord with the results in Figure 7, which show that a long time is spent in the entropic basin in nearly all transitions. However, in most cases, the trajectories go through the native basin and can be diagrammed as ($i \rightarrow$ entropic $\rightarrow$ native $\rightarrow$ entropic $\rightarrow j$), often spending a long time in the native basin and making repeated transitions ($i \leftrightarrow$ entropic $\leftrightarrow$ native $\leftrightarrow$ entropic $\leftrightarrow j$). There is essentially full equilibration in the native basin with rapid sampling of different conformations; there are 2672 nodes (different secondary structure strings) in the native basin. Further, if one examines the content of native secondary structure in the last 0.2 ns before the trajectory exits from the folded to the entropic state, with the condition that it will next visit a certain enthalpic basin (e.g., Ns-or or Cs-or), a significant structural bias toward that basin is already present (see Figure 8). Thus, the fate of the trajectory is biased already in the native state. This analysis is in accord with the conclusion that the system does not stay in the entropic state long enough to equilibrate. The latter is due in part to the aforementioned low barrier between the entropic state and the native state (see the left-most profile of Figure 7) and in part to the presence of significant barriers between different parts of the entropic basin. This is shown in section E of the Supporting Information by the reduced pfoldf FEPs of pairs of non-native basins and the entropic state (e.g., Ns-or, entropic, and Cs-or in Figure S6a). However, not all basins of the denatured state are separated by barriers within the entropic state (one example is in Figure S6c for the Ns-or and the cyan basins). These observations explain the origin of the barriers in the entropic region at $0.3 \leq Z_A \leq 0.4$ for the reduced FEPs (i.e., FEPs calculated taking into account only one basin and the entropic and native states) from the Cs-or or the helical basin in Figure 7. In other words, Figures 7 and S6 are consistent because both show barriers in the entropic region mainly between Cs-or, Ns-or, and helical regions, but not among conformations with the C-terminal hairpin folded (i.e., Ns-or, Ch-curl$_1$, Ch-curl$_2$, cyan, and magenta basins in Figure 7).

The number of folding transitions from non-native basins with a structured C-terminal hairpin is larger than the corresponding number from basins with the N-terminal hairpin formed (Cs-or). This observation is consistent with the analysis of folding transition state structures[19] identified by a node- $p_{fold} = 0.5$ criterion.[34] The two main folding pathways of Beta3s are in agreement with the diffusion-collision model[52] in which the folding process involves the encounter of marginally stable secondary structural elements.[53] The SEKN also sheds light on

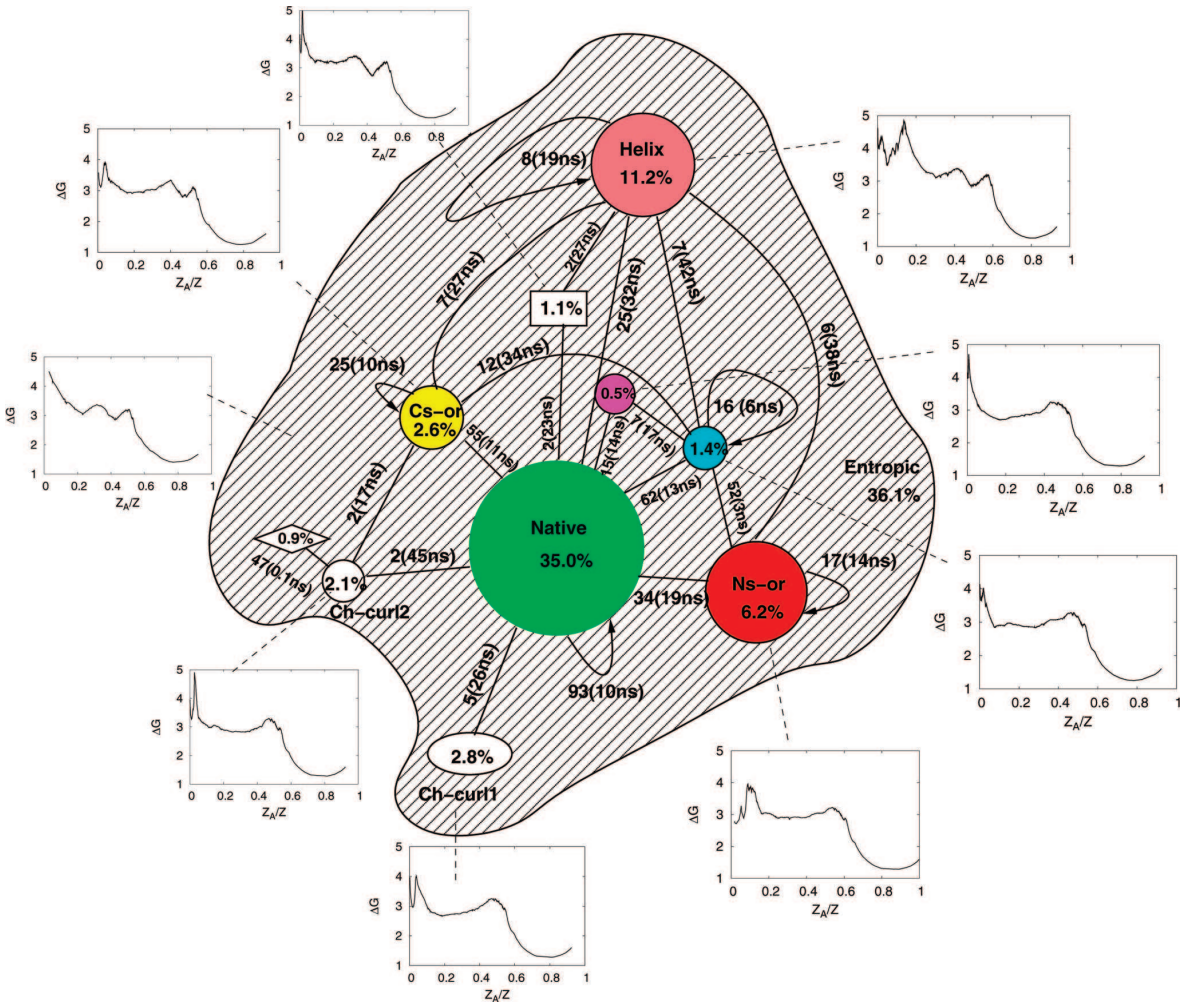


Figure 7. Simplified equilibrium kinetic network (SEKN) of Beta3s free-energy basins. The circles, ellipse, rectangle, and diamond are the 10 most populated free-energy basins of Beta3s with their respective statistical weight as isolated by pfoldf. Colors and geometrical shapes were chosen to be consistent with the basins identified by KGA.[19] The dashed surface represents the entropic basin that is made up of all conformations not belonging to any of the 10 basins. Plotting the entropic basin as a circle would be incorrect given that there is no fast equilibration inside it because of the low barrier toward the native state. Each link shows the number of transitions between the bottom (most populated secondary structure string) of basin *i* and the bottom of basin *j* through the entropic basin only (i.e., without visiting any other basin bottom), and the average time spent in the entropic state is given in parentheses. Transitions from non-native enthalpic basins and from the entropic state to the native state are illustrated by pfoldf FEPs, where the representative node of the entropic state is the extra node. Each profile from a non-native basin was calculated using only the basin under consideration (dashed line), the entropic state, and the native state while neglecting the remaining eight basins. The profile from the entropic state (left middle) takes into account only the entropic and native states. Each profile gives a description of both the entropic/native barrier ($Z_A/Z \approx 0.5$) and the barrier to leave the enthalpic basin ($Z_A/Z \lesssim 0.1$), and these regions are both devoid of overlap.

the folding pathways from the helical state (Figure 7). Half of the transitions proceed via the entropic state, where the system spends considerable time, directly to the folded state. Less frequently, the trajectory visits Cs-or or Ns-or structures or even returns to the helical basin. We note that, even though the folding times from the Ns-or and Cs-or conformations are relatively large (110 and 70 ns, respectively) and the Ns-or and Cs-or conformations hardly interconvert, the difference between the most populated nodes of the three basins is only in one to three positions of the secondary structure string. This observation illustrates that a small structural change can result in a large kinetic distance. For instance, the structural change from folded to Cs-or, i.e., from two to three S letters at the second turn (see Table 2 for strings), involves a complete rearrangement of the side chains of the C-terminal strand.

For each of the basins identified in the SEKN, the analysis showed that the equilibration within the basin is fast, relative to transitions from or to the basin. However, this is not true for the entropic basin. The low barrier between the entropic state and the native state together with the high population of the former (about 33%) leads to fast transitions to the native state that prevent equilibration in the entropic basin. Therefore, plotting the entropic state as a single node would not only be misleading, but would also give an incorrect picture of the pathways. If the entropic state were replaced by one node, the native state would be connected to only the entropic node, and the picture would suggest that the entropic state acts as the hub. However, in the SEKN emerging from the pfoldf procedure, both the native and the entropic states can be considered to be hubs.

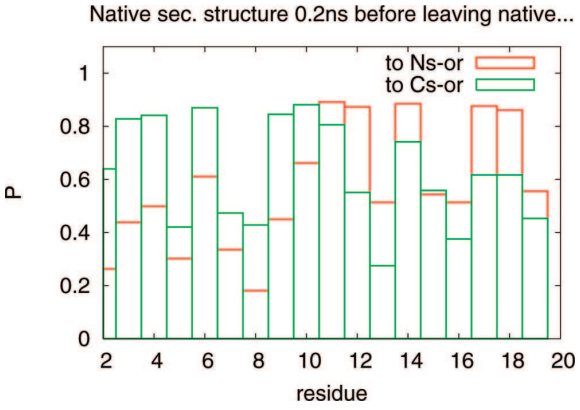


Figure 8. Native secondary structure content of the trajectory in the folded state, 10 snapshots (0.2 ns) before the system leaves toward one of the two non-native enthalpic basins Ns-or or Cs-or. Notably, if the trajectory continues to Ns-or (Cs-or), the N-terminal (C-terminal) hairpin is already disrupted before leaving the native state.

Transition Disconnectivity Graph (TRDG). The TRDG[44] of Beta3s (Figure 9) provides further evidence that the denatured state is heterogeneous and has several funnel-like basins with favorable effective energies as well as a helical basin. Disconnectivity graphs do not visualize the basins that lack representative nodes, i.e., the basins with high entropic contributions to the free energy. The split of the helical basin is consistent with the FEPs of Figures 2 and 3 and explains the longer commitment time required in KGA to isolate the complete helical basin, as discussed in the Results section. An advantage of TRDG over one-dimensional FEPs and the SEKN is that it quantitatively depicts (mainly enthalpic) minima and barriers in a single plot.

Single-Exponential Kinetics of Folding. The cumulative distribution of folding times of all configurations in the trajectory shows single-exponential behavior (Figure 10 top). This (apparently simple) behavior is consistent with the fact that the barriers to exit the individual basins are of similar heights in the FEPs from enthalpic traps or the helical basin (Figure 3 and Table 3). Together with the aforementioned low barrier from the entropic state to the native state, the similar barrier heights explain the single-exponential behavior. The similar free-energy barriers from Ns-or and Cs-or are likely to be a consequence of the high sequence identity (67%) between the N-terminal β-hairpin (residues 1−12) and the C-terminal β-hairpin (residues 9−20). On the other hand, there is no straightforward explanation for the similar barrier height from the helical basin.

Model rate calculations for a photoswitchable peptide have shown that, for an SEKN similar to the present one, a spread in the rates on the order of a factor of 9 is required to observe significant deviations from single-exponential behavior.[54]

Because the entropic state cannot be represented as a single node in the SEKN (because of slow relaxation), the distribution of folding times was obtained by simulating MC kinetics on the network rather than on the SEKN. An initial point was picked arbitrarily among nodes that do not belong to the native basin with a probability proportional to the statistical weight, and MC simulations were performed until the system reached a node corresponding to the native basin. Figure 10 (bottom) shows the density function of the distribution obtained with 10^5 trajectories and a single-exponential distribution with the corresponding folding time. The curves are in reasonable agreement, which indicates that the kinetics can be approximately described as single-exponential.

Interestingly, the present analysis indicates that somewhat deceptive single-exponential behavior can emerge from completely different energy landscapes as for both Beta3s, which has a kinetically partitioned denatured state and a hub-like native state, and the β-hairpin of protein G, which shows fast equilibration within a multibasin denatured state.[22]

Concluding Discussion

Considerable progress has been made recently in experimental investigations of protein folding. Particularly, for the small fast-folding proteins, which have been studied most,[4] a two-state description (folded and unfolded) is adequate to describe the measurements. This means that little, if any, information concerning the details of the folding pathways is obtained, although mutation studies have been used to provide a coarse-grained description of the transition state.[55,56] Also, expirements supplementing the kinetic measurements by probes sensitive to structural details[6,54] have yielded some insights into the folding pathways, particularly when intermediates are present.[57] However, none of the experimental studies provide a detailed description of the structures that contribute significantly to the ensembles that are sampled along the folding pathways. Although one might hope for such information from future experiments, as of now, the only way to approach this problem is by computer simulations. In recent years, aided by faster, often massively parallel computational resources, the first steps toward this aim have been realized. Mostly, the studies have been limited to peptides[22,32,42] and a few miniproteins (e.g., Trp cage) that fold on the simulation (nanosecond) time scale. Unfolding simulations at high temperature have been interpreted in the folding direction,[58] but with few exceptions,[59] the unfolding reaction has been followed only once in the stimulations. There are very few systems for which it has been possible to do the multiple folding simulations required to obtain statistically meaningful results for analysis. The Beta3s miniprotein, with an implicit solvent model, is one such system. The present analysis is based on equilibrium simulations 20 μs in length that show about 100 folding/unfolding events for a temperature (330 K) at which the native and denatured states are both significantly populated (35% native and 65% denatured).

One problem in using the simulation results is the difficulty of analyzing them to obtain an understanding of the folding reaction. The number of degrees of freedom (3×215 for a small system such as Beta3s in the polar hydrogen approximation, in which aliphatic and aromatic hydrogen atoms are not considered explicitly) makes a straightforward approach impossible, even though all of the details (hopefully representative of the actual folding process) are available from the trajectory. Use of the results requires a method for reducing the problem to one or only a few dimensions that are sufficient to describe the folding reaction in a meaningful way. The recently developed cut-based FEP procedures and complex network analyses have been shown to essentially solve this problem. They have demonstrated, among other conclusions, that the very simple picture of protein folding (e.g., one or at most two barriers between the denatured and native state), often obtained by projecting the free energy on an arbitrarily chosen progress variable(s), is not consistent with the complexity of the actual free-energy surface.[18,22,32,42] Such complexity has spurred the development of more sophisticated computational procedures for determining free-energy basins and transitions among them. The essential element of both the p_{fold}-/mfpt-based procedures for FEP calculation[24] and kinetic grouping analysis[19] is the

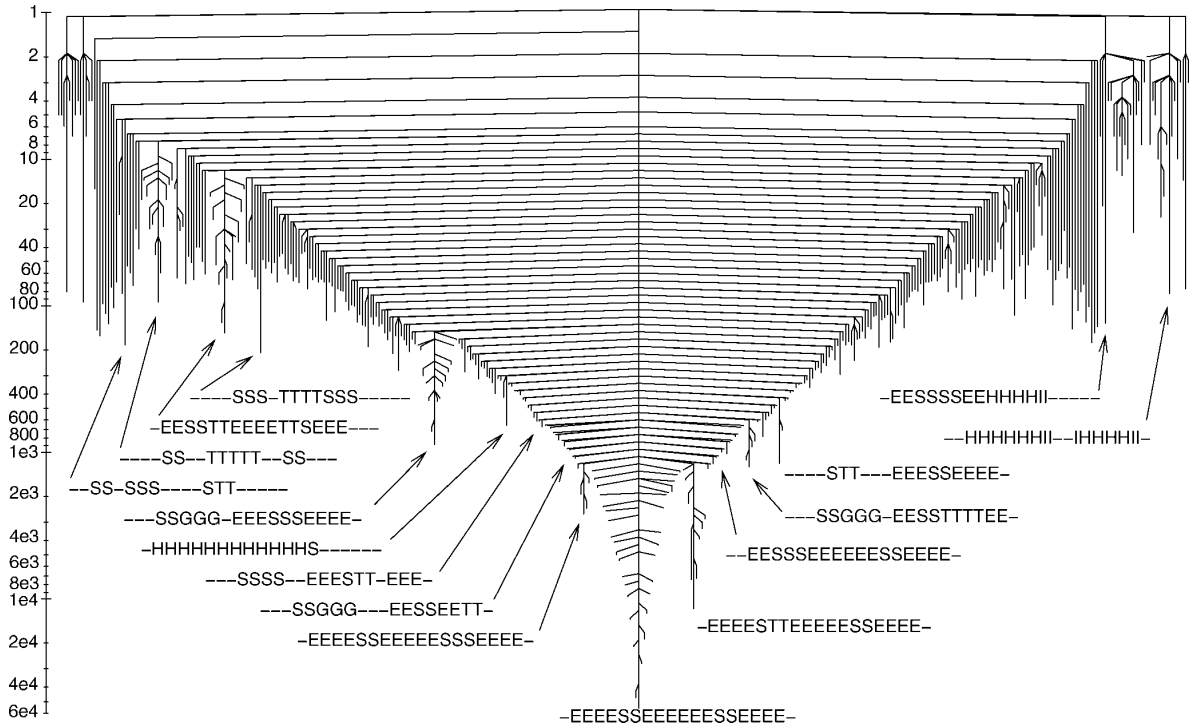


Figure 9. Free-energy (transition) disconnectivity graph of Beta3s EKN obtained with secondary structure coarse-graining. Secondary structure strings of the most populated clusters in the basins are shown. The vertical axis shows the number of times the secondary structures were visited (the bottom of each vertical line) and the number of transitions between pairs of strings (the value where the lines corresponding to two strings intersect). Information about the strings is given in Table 2.

identification of free-energy basins, not according to geometrical characteristics (such as the fraction of native contacts or rmsd from the folded structure) but rather according to the transitions that occur in folding/unfolding trajectories at equilibrium. From such an analysis, a meaningful one-dimensional projection of the free-energy surface (called FEP in this work) is obtained. It provides the basins on the surface and the barriers between them. Unlike the standard projections, which lead to overlap of the basins that smooth out the barriers and can make them disappear (as shown for the β-hairpin of protein G in ref 24), the progress variables used here do not result in such overlap if a valid clustering algorithm is used. The formulation requires a coarse-graining approach to group snapshots saved along equilibrium trajectories into nodes, so that adequate transition statistics can be obtained. Some evidence for the robustness upon changing the coarse-graining algorithm (based on rmsd or secondary structure string) was provided previously[22,32] and is given in the Supporting Information, but this issue should be analyzed in more detail.

An important conclusion from the present study is that the p_{fold}-/mfpt-based procedures[24] find the same free-energy basins as the kinetic grouping approach described in a previous work.[19] The very similar results for the free-energy surface, as well as the identification of the subtle differences between Beta3s and its W10V mutant, indicate that the two approaches are correct and complementary. One aspect of the p_{fold}-/mfpt-based procedures is that they are able to separate a free-energy basin consisting of an ensemble of conformations with fluctuating helical (i.e., non-native secondary structure) content. The helical basin has a population of about 11%, and from this basin, the folded state is reached through a heterogeneous entropic state. The significant statistical weight of the helical basin is surprising if one considers that Beta3s is a peptide designed to assume a three-stranded antiparallel β-sheet fold. Because of its entropic stabilization, the helical state of Beta3s is expected to be less populated at lower temperature. Interestingly, an α-helix-rich kinetic intermediate in the refolding (by chemical denaturant dilution) of the β-sandwich protein src SH3 has been reported on the basis of circular dichroism, fluorescence, and X-ray solution scattering experiments.[60] Moreover, at pH 3, a helical equilibrium intermediate of the A45G mutant of src SH3 has been observed, and evidence has been provided that it corresponds to a kinetic intermediate.[61]

The differences between Beta3s and its W10V mutant are small but relevant. There is a shift of the equilibrium in the helix/β-sheet statistical-weight ratio from 11/35 for the wild type to 9/40 for W10V. This indicates that even a single-point mutation can have an influence on the relative propensity of secondary structure formation that plays a critical role in diseases related to protein misfolding and aggregation.[62-64]

About one-third of the snapshots saved along the molecular dynamics trajectories belong to a heterogeneous entropic state that is visited during individual transitions between mainly enthalpic free-energy basins. There is no fast equilibration within the entropic state because of the low barrier of only around 0.5 kcal/mol toward the native state, but also because of barriers that split the entropic state. This explains the previously reported kinetically partitioned denatured state.[19] As a consequence, the future development of the system depends strongly on the region where the system enters the entropic state or, equivalently, where it exits the native basin (Figure 8). Because the native and entropic states together make up almost 70% of the total weight, the free-energy surface of Beta3s seems to exhibit some of the features of a fast barrierless/low-barrier folder but with meta-

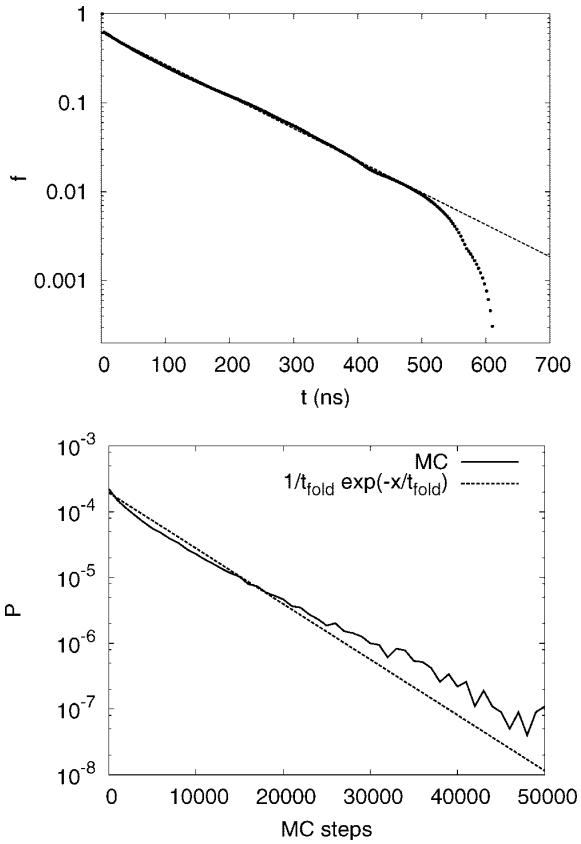


Figure 10. Single-exponential behavior of folding. (Top) Points show the cumulative distribution of the first passage time to the folded state $f(t) = \int_t^\infty p(\tau)\, d\tau$, where p is the probability distribution of the first passage time. All saved snapshots were used to calculate $f(t)$. The dashed line is the exponential fit $0.624e^{(-t/120)}$. (Bottom) Distribution of first passage times obtained from MC simulations on the directed network of 2.5-Å rmsd coarse-graining (solid line). The single-exponential approximation is also shown (dashed) with $t_{fold} = 5140$ MC steps, which corresponds to a folding time of 102.8 ns (5140 × 0.02 ns). The 2.5-Å rmsd clustering was used here because the corresponding distribution calculated with the secondary structure coarse-graining is affected by pseudotunneling (see text and Figure S3 in the Supporting Information).

stable enthalpic traps, each with a relatively low population and a total weight of about 20%, plus a helical region populated at about 11%.

The FEPs reveal that the barriers to exit the enthalpic traps have similar heights, and the SEKN shows that the times spent in the entropic state before reaching or after leaving the folded state are comparable. These results explain why the folding times from individual basins differ by no more than a factor of 3.[19] Thus, in accord with a recent experimental analysis of a photoswitchable helical peptide,[54] the single-exponential folding behavior originates from esentially equal folding times for multiple paths. This provides another scenario, different from that of the β-hairpin of protein G,[22] by which the complexity of the folding reaction can be hidden from standard kinetic experiments.

Acknowledgment. We thank F. Rao, E. Guarnera, and E. Paci for contributions to the initial stages of this work. We thank M. Seeber for help with the program WORDOM. The simulations were performed on the Matterhorn cluster of the University of Zurich. This work was supported by a Swiss National Science Foundation grant to A.C., and the portion done at Harvard University was supported, in part, by a grant from the National Institutes of Health. S.K. was supported by the CHARMM Development Project. Procedures for calculating the FEPs have been introduced into WORDOM (http://www.biochem-caflisch.unizh.ch/wordom), and the program for plotting the TRDG (Figure 9) is available upon request.

Supporting Information Available: Additional calculations, figures, and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 11–27.
(2) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
(3) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1993**, *232*, 660–679.
(4) Maxwell, K. L.; Wildes, D.; Zarrine-Afsar, A.; De Los Rios, M. A.; Brown, A. G.; Friel, C. T.; Hedberg, L.; Horng, D.; Bona, J.-C.; Miller, E. J.; Vallée-Bélisle, A.; Main, E. R. G.; Bemporad, F.; Qiu, L.; Teilum, K.; Vu, N.-D.; Edwards, A. M.; Ruczinski, I.; Poulsen, F. M.; Kragelund, B. B.; Michnick, S. W.; Chiti, F.; Bai, Y.; Hagen, S. J.; Serrano, L.; Oliveberg, M.; Raleigh, D. P.; Wittung-Stafshede, P.; Radford, S. E.; Jackson, S. E.; Sosnick, T. R.; Marqusee, S.; Davidson, A. R.; Plaxco, K. W. *Protein Sci.* **2005**, *14*, 602–616.
(5) Jackson, S. E. *Fold. Des.* **1998**, *3*, R81−91.
(6) Dobson, C. M.; Šali, A.; Karplus, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 869–893.
(7) Dinner, A. R.; Šali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331–339.
(8) Mirny, L. A.; Shakhnovich, E. I. *Ann. Rev. Biophys. Biomolec. Struc.* **2001**, *30*, 361–396.
(9) Daggett, V.; Fersht, A. R. *Nature Rev. Mol. Cell Biol.* **2003**, *4*, 497–502.
(10) Wolynes, P. G. *Phil. Trans. R. Soc. A* **2005**, *363*, 453–467.
(11) Chan, H. S.; Dill, K. A. *Proteins: Structure, Function, and Bioinformatics* **1998**, *30*, 2–33.
(12) Best, R.; Hummer, G. *Proc. Natl. Acad. Sci. USA.* **2005**, *102*, 6732–6737.
(13) Palyanov, A. Y.; Krivov, S. V.; Karplus, M.; Chekmarev, S. F. *J. Phys. Chem. B* **2007**, *111*, 2675–2687.
(14) Go, N.; Abe, H. *Biopolymers* **1981**, *20*, 991–1011.
(15) Schueler-Furman, O.; Wang, Ch.; Bradley, P.; Misura, K.; Baker, D. *Science* **2005**, *310*, 638–642.
(16) Kussell, E.; Shimada, J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. USA.* **2002**, *99*, 5343–5348.
(17) Paci, E.; Vendruscolo, M.; Karplus, M. *Proteins: Structure, Function, and Bioinformatics* **2002**, *47*, 379–392.
(18) Caflisch, A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 71–78.
(19) Muff, S.; Caflisch, A. *Proteins: Structure, Function, and Bioinformatics* **2008**, *70*, 1185–1195.
(20) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495–1517.
(21) Wales, D. J. *Energy Landscapes*, Cambridge Univ. Press, Cambridge, U.K., 2003.
(22) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. USA.* **2004**, *101*, 14766–14770.
(23) Munoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.
(24) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(25) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. I. *J. Chem. Phys.* **1998**, *108*, 334–350.
(26) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. USA.* **2000**, *97*, 10780–10785.
(27) De Alba, E.; Santoro, J.; Rico, M.; Jiménez, M. A. *Protein Sci.* **1999**, *8*, 854–865.
(28) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035–4042.
(29) Cavalli, A.; Ferrara, P.; Caflisch, A. *Proteins: Structure, Function, and Bioinformatics* **2002**, *47*, 305–314.
(30) Serrano, L.; Matouschek, A.; Fersht, A. R. *J. Mol. Biol.* **1992**, *224*, 805–818.
(31) Settanni, G.; Rao, F.; Caflisch, A. *Proc. Natl. Acad. Sci. USA.* **2005**, *102*, 628–633.
(32) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(33) Hubner, I. A.; Shimada, J.; Shakhnovich, E. I. *J. Mol. Biol.* **2004**, *336*, 745–761.

(34) Rao, F.; Settanni, G.; Guarnera, E.; Caflisch, A. *J. Chem. Phys.* **2005**, *122*, 184901.
(35) Snow, C. D.; M Rhee, Y.; Pande, V. S. *Biophys. J.* **2006**, *91*, 14–24.
(36) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
(37) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.
(38) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.
(39) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Structure, Function, and Bioinformatics* **2002**, *46*, 24–33.
(40) Ferrara, P.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. B* **2000**, *104*, 5000–5010.
(41) Eaton, W. A.; Munoz, V.; Hagen, J.; Jas, S. G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. *Ann. Rev. Biophys. Biomolec. Struc.* **2000**, *29*, 327–359.
(42) Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. USA.* **2006**, *103*, 17747–17752.
(43) Andersen, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. *Structure* **2002**, *10*, 174–184.
(44) Krivov, S. V.; Karplus, M. *J. Chem. Phys.* **2002**, *117*, 10894–10903.
(45) Park, S.; Sener, M. K.; Lu, D.; Schulten, K. *J. Phys. Chem.* **2003**, *119*, 1313–1319.
(46) Apaydin, M.; Brutlag, D.; Guesttin, C.; Hsu, D.; Latombe, J. *In International Conference on Computational Molecular Biology (RECOMB)* **2002**,
(47) Ford, L. R.; Fulkerson, D. R. *Canadian J. of Math.* **1956**, *8*, 399–404.
(48) Gomory, R. E.; Hu, T. C. *SIAM J. Applied Math.* **1961**, *9*, 551–570.
(49) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Structure, Function, and Bioinformatics* **1995**, *21*, 167–195.
(50) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Munoz, V. *Science* **2002**, *298*, 2191–2195.
(51) Eaton, W. A.; Munoz, V.; Thompson, P. A.; Chan, C. K.; Hofrichter, J. *Curr. Opin. Struct. Biol.* **1997**, *7*, 10–14.
(52) Karplus, M.; Weaver, D. L. *Biopolymers* **1979**, *18*, 1421–1437.
(53) Pappu, R. V.; Weaver, D. D. *Protein Sci.* **1998**, *7* (2)), 480–490.
(54) Ihalainen, J. A.; Bredenbeck, J.; Pfister, R.; Helbing, J.; Woolley, G. A.; Hamm, P. *Proc. Natl. Acad. Sci. USA.* **2007**, *104*, 5383–5388.
(55) Matouschek, A.; Kellis, J. T., Jr.; Serrano, L.; Fersht, A. R. *Nature* **1989**, *340*, 122–126.
(56) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *Nature* **2001**, *409*, 641–645.
(57) Radford, S. E.; Dobson, C. M.; Evans, P. A. *Nature* **1992**, *358*, 302–307.
(58) Fersht, A. R.; Daggett, V. *Cell* **2002**, *108*, 573–582.
(59) Day, R.; Daggett, V. *Proc. Natl. Acad. Sci. USA.* **2005**, *102*, 13445–13450.
(60) Li, J.; Shinjo, M.; Matsumura, Y.; Morita, M.; Baker, D.; Ikeguchi, M.; Kihara, H. *Biochemistry* **2007**, *46*, 5072–5082.
(61) Li, J.; Matsumura, Y.; Shinjo, M.; Kojima, M.; Kihara, H. *J. Mol. Biol.* **2007**, 747–755.
(62) Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M. *Nature* **2003**, *424*, 805–808.
(63) Tartaglia, G. G.; Cavalli, A.; Pellarin, R.; Caflisch, A. *Protein Sci.* **2004**, *13*, 1939–1941.
(64) Tartaglia, G. G.; Cavalli, A.; Pellarin, R.; Caflisch, A. *Protein Sci.* **2005**, *14*, 2723–2734.

JP711864R

One-Dimensional Barrier Preserving Free-Energy Projections of a β-sheet Miniprotein: New Insights into the Folding Process SUPPLEMENTARY MATERIAL

Sergei Krivov,1† Stefanie Muff,2† Amedeo Caflisch,2* and Martin Karplus,1,3*

1 Laboratoire de Chimie Biophysique, ISIS F-67000 Strasbourg, France

2 Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

and

3 Department of Chemistry & Chemical Biology Harvard University Cambridge, Massachusetts 02138 U.S.A. †

Keywords:

A. Iterative calculation of $p_{fold}(\tau_{commit})$ on the EKN

The calculation of p_{fold} values basing on the equilibrium kinetic network (EKN) has been described previously[1] and in the main text. The calculation of $p_{fold}(\tau_{commit})$ values is based on a different system of equations and therefore requires additional considerations. Let p_i be the p_{fold} of node i. Then:

$$p_i = P[\tau_f(i) \leq \tau_{commit}] \ ,$$

with $\tau_f(i)$ representing the first passage time to the native node, starting in node i. Given a simulation with saving frequency Δt, the system of equations to be solved is

$$\begin{aligned} P[\tau_f(i) \leq \tau_{commit}] &= \sum_j p_{ji} P[\tau_f(j) \leq \tau_{commit} - \Delta t] \\ &= \sum_j p_{ji} \left(P[\tau_f(j) \leq \tau_{commit}] - P[\tau_f(j) = \tau_{commit}] \right) \ , \end{aligned} \tag{1}$$

where p_{ji} is the transition probability from i to j and the sum runs over all nodes of the EKN. The system is bound by the condition $p_A = 1$. Let us first evaluate $P[\tau(j) = k], \Delta t \leq k \leq \tau_{commit}$, where $P[\tau(j) = k] = P[T_k = A | T_0 = j]$ with T_K equal to the probability to be in the native node A after k steps, starting from node j (not necessarily the first passage time). To avoid costly multiplication of the whole transition matrix, it is easier to evaluate the "reverse" probability to be in node j after k steps starting in A, $P[T_k = j | T_0 = A]$, because this can be calculated at once by iterative multiplication of the starting configuration $P[T_0 = j | T_0 = A] = \delta_{j,A}$ by the transition matrix:

$$P[T_{k+\Delta t} = j | T_0 = A] = \sum_i p_{ji} P[T_k = i | T_0 = A] \ .$$

Since the EKN fulfills detailed balance, the probability of the $j \rightarrow A$ transition can be calculated by

$$\underbrace{P[T_k = A | T_0 = j]}_{= P[\tau(j) = k]} = P[T_k = j | T_0 = A] \cdot \frac{P[A]}{P[j]} \ ,$$

where $P[A], P[j]$ are the relative populations of the nodes. The probability for the *first* passage time τ_f to node A can thus be calculated by

$$P[\tau_f(j) = \tau_{commit}] = \left(\prod_{n=1}^{(\tau_{commit} - \Delta t)/\Delta t} (1 - P[\tau(j) = n\Delta t]) \right) \cdot P[\tau(j) = \tau_{commit}] \ ,$$

i.e., the probability *not* to return within $\tau_{commit} - \Delta t$, but within exactly τ_{commit}. Inserting this expression into equation (1) and solving the system of equations yields the correct folding probabilities.

Figure S1 shows the FEP of Beta3s for different values of τ_{commit} and makes clear that too short commitment times are not suitable to fully resolve the unfolded state.

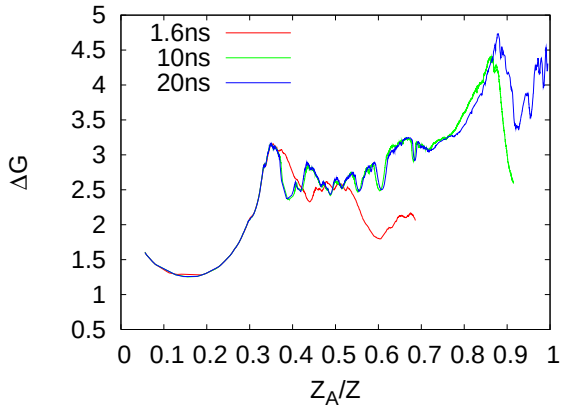


FIG. S1: FEP with $p_{fold}(\tau_{commit})$ calculations for τ_{commit}=1.6ns, 10ns and 20ns. It is important to choose the commitment time long enough to resolve the unfolded state. In fact, using a τ_{commit} value of only 1.6 ns (red curve) about 30% of the conformations have $p_{fold} = 0$ so that the profile stops at Z_A/Z=0.7.

B. Differences between pfoldf and mfpt FEPs

The deviations between the FEPs obtained by the two procedures originate from at least two points. First, p_{fold} calculations are bound by two conditions (p_A=1, p_B=0), and mfpt calculations only by one (τ_A=0). Second, the p_{fold} values are calculated on a slightly different (biased) underlying EKN due to the extra node that is used in the pfoldf procedure. When the nodes are sorted according to decreasing p_{fold} or increasing mfpt, the Spearman correlation coefficient of the noderanks (ρ) decreases with increasing λ (for λ=0.0001: ρ=0.9997; for λ=0.01: ρ=0.988), because a larger λ enhances the bias. If the mfpt and pfoldf FEPs are calculated on the same underlying EKN with the extra node connected with capacity λ=0.0001, pfoldf and mfpt are still not identical, although

very similar with ρ=0.9999.

C. Mfpt as progress coordinate

The progress coordinate of the FEPs is the relative partition function of the EKN Z_A/Z, so that no information on the the underlying progress variable (p_{fold}, $p_{fold}(\tau_{commit})$ and mfpt) is present in the final plot. It is, however, straightforward to project the profile onto the original variable. In this way, the progress coordinate and the underlying progress variable are the same. Such a transformed profile shows ΔG as a function of the kinetic distance (in time units) from the native state (Figure S2) and provides supplementary information to the Z_A/Z projection. A disadvantage of the projection onto mfpt is that the non-native enthalpic basins are very close together in the profile because most of them have similar mfpt values (especially on the secondary structure network, where most values are around 10 ns for the mfpt values calculated by numerical solution of the equation $\mathrm{mfpt}_i = \Delta t + \sum p_{ji} \cdot \mathrm{mfpt}_j$, as detailed in the Methods section of the main text). Note that for the network with nodes coarse-grained according to secondary structure the numerically calculated mfpt values are smaller than those calculated directly from the trajectory (i.e., if one would follow the trajectory each time a node is visited), which arises from the fact that the secondary structure coarse-graining is too generous (see below) and because the solution of the mfpt equation system is equivalent to running a very long (infinite) Monte-Carlo (MC) simulation. On the other hand, performing the same analysis on the network obtained by the 2.5 Å RMSD coarse-graining, the mfpt values are very close to those found directly from the trajectory.

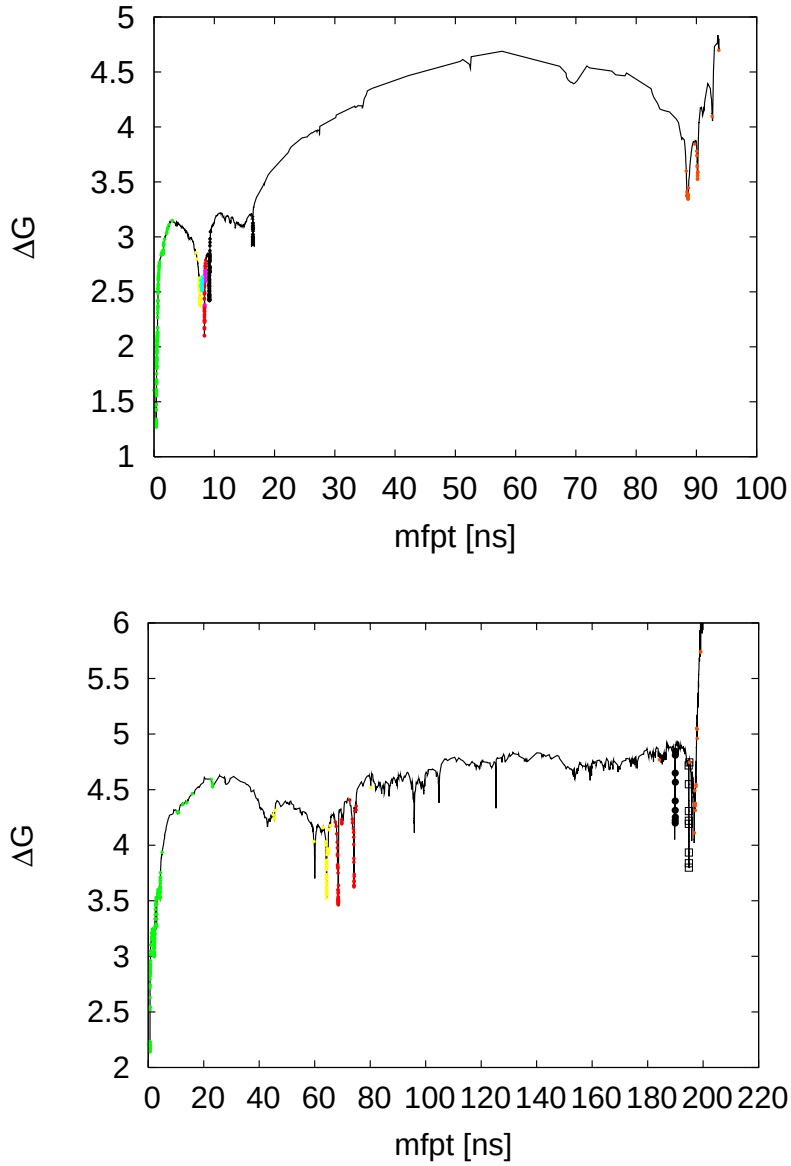


FIG. S2: Beta3s unfolding FEP calculated for the EKN (see Methods) using mfpt as a progress coordinate *and* progress variable for the secondary structure (top) and 2.5 Å RMSD coarse-graining (bottom). As in Figure 4 of the main text, individual basins are colored according to the basins extracted by the pfoldf procedure. Note that values of mfpt for individual basins are larger, and the barrier separating the native basin from the rest is higher for 2.5 Å RMSD than for secondary structure coarse-graining because of pseudo-tunneling affecting mainly the latter. Importantly, the similar mfpt values for the enthalpic traps, and a spread of only about three between mfpt values of enthalpic traps and the helical basin, is consistent with the single-exponential behavior of folding (see Results section in the main text).

D. Coarse-graining and Monte Carlo simulations

The main disadvantage of RMSD coarse-graining (clustering is used here as a synonymous) is the required computer time. For the one million snapshots of the 20 μs trajectory, all-atom RMSD clustering with a cutoff of 2.5 Å and 2.0 Å requires 10 days and (an estimate) 40 days, respectively, on a 2.8 GHz Intel Xeon. On the other hand, the disadvantage of secondary structure coarse-graining is revealed if MC simulations are performed on the resulting directed network. The folding time decreases from 100 ns to about 10 ns, whereas MC simulations on the directed network obtained by all-atom RMSD coarse-graining with 2.5 Å cutoff yield the correct value of 100 ns. Interestingly, a finer graining (RMSD 2.0 Å) increases the folding time to 137 ns, whereas the coarser RMSD of 3.0 Å decreases it to 84 ns. This phenomenon can be explained as follows: A very fine grained clustering yields low populations even for clusters in the native state. With a non-neglectable probability it can then happen that a folding event is not accounted because the trajectory does not visit the most populated (native) node before it unfolds, because the cluster is too small. On the other hand, a very coarse assignment of nodes as for RMSD 3.0 Å or secondary structure reduces the folding time in the MC simulation. The reason for the latter is that, due to the lax restriction of nodelimits, pseudo-tunneling happens frequently between nodes that are in reality separated by a significant barrier. Each pseudo-tunneling event introduces a "shortcut" into the network which is taken into account in the MC simulation, even though folding never really proceeds via such a shortcut in the molecular dynamics (MD) simulation. This property leads to non-Markovianity. It has been observed earlier that the equivalence between MC and MD kinetics does not follow automatically and depends on the coarse-graining procedure[2] (Figure S3).

Interestingly, despite the considerable differences between the two methods used for coarse-graining, the basins isolated by pfoldf with secondary structure or 2.5 Å RMSD clustering are almost identical (Table S-I). Each snapshot belongs to a coarse-grained conformation, so it is grouped to the basin of the respective conformation. Basins are therefore comparable snapshot by snapshot and a similarity can be calculated analogous to Table II in the main text. Both the KGA and pfoldf procedures are not noticeably affected by the shortcuts (i.e., by the non-Markovian character) in the secondary structure

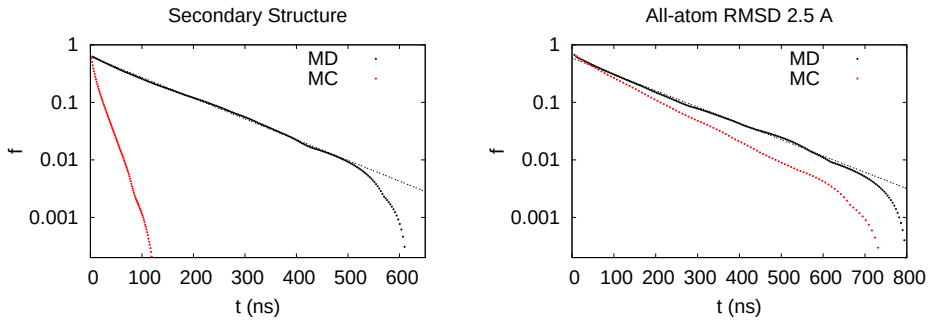


FIG. S3: Cumulative distribution of first passage times for the secondary structure (left) and the 2.5 Å RMSD coarse-graining (right). Black dots are extracted from the MD simulation, red dots from a 200μs MC simulation on the directed network. The folding kinetics of the secondary structure-based MC trajectory differ considerably from MD kinetics, whereas with 2.5 Å RMSD clustering almost identical MC and MD folding kinetics are observed.

coarse-graining. However, the free-energy barrier in the unfolding FEP of the native state (obtained by pfoldf) is about 0.5 kcal/mol higher using the all-atom 2.5 Å RMSD clustering (Figure S4) than the secondary structure clustering (Figure 2 of the main text), which shows that, in contrast to the isolation of basins, the extraction of barriers is very sensitive on the coarse-graining.

Basin		Weight (%)		Number of nodes		
Heaviest node	Name	sstruct	RMSD	sstruct	RMSD	Similarity[a]
-EEEESSEEEEEESSEEEE-	Native	35.0	36.4	2672	6457	99.5
-EEEESTTEEEEESSEEEE-	Ns-or	6.2	3.2/2.9	1278	220/798	98.4/95.8
-EEEESSEEEEESSSEEEE-	Cs-or	2.6	3.8	967	5167	98.5
-HHHHHHHHHHHHS------	Helix	11.6	11.2	57134	49049	95.4
---SSGGG---EESSEETT-	Ch-curl$_1$	2.8	2.8	2153	430	95.0
---SSGGG-EESSTTTTEE-	Ch-curl$_2$	2.1	2.0	1675	119	98.8

TABLE S-I: Comparison of most populated basins of Beta3s obtained by pfoldf using either secondary structure or all-atom 2.5 Å RMSD clustering. Ns-or is split into two basins of almost equal size for RMSD, but the partitioning is also visible in the one-dimensional FEP generated using secondary structure clustering (Figure 3 of the main text). [a]The similarity value is calculated as the intersection of two corresponding basins, normalized to the lower population.

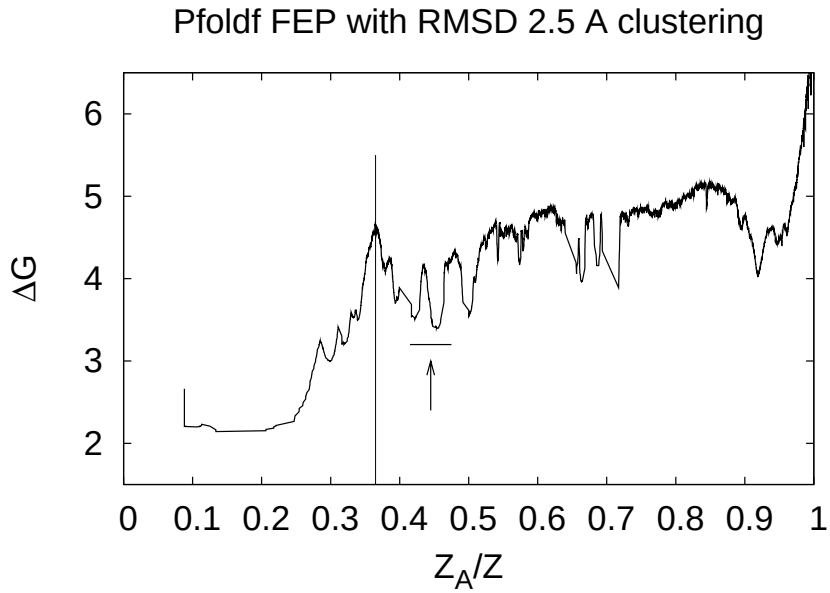


FIG. S4: Pfoldf-FEP of Beta3s using the snapshots (from the MD trajectories) clustered according to all-atom 2.5 Å RMSD. The vertical line shows the position of the unfolding barrier as extracted from the pfoldf procedure. The arrow and horizontal segment indicate the Ns-or basin which is split into two using 2.5 Å RMSD clustering. Note that this profile is very similar to the one obtained using secondary structure coarse-graining (Figure 2 top of the main text), but the barrier of the native basin is higher for 2.5 Å RMSD. In both the most distant basin from native is the helical basin.

A main difference between secondary structure and all-atom RMSD coarse-graining is that the former lacks the information about the position and orientation of the sidechains. Therefore, it is possible that conformations belonging to the same secondary structure string are separated by barriers that arise from differences in the orientation of sidechains. To exemplify the concern, Figure S5 shows two structures belonging to the native secondary structure node (`-EEEESSEEEEEESSEEEE-`), one with the Tyrosine19 sidechain pointing upward and one pointing down. The 2.5 Å RMSD coarse-graining correctly separates these two structures into two different clusters.

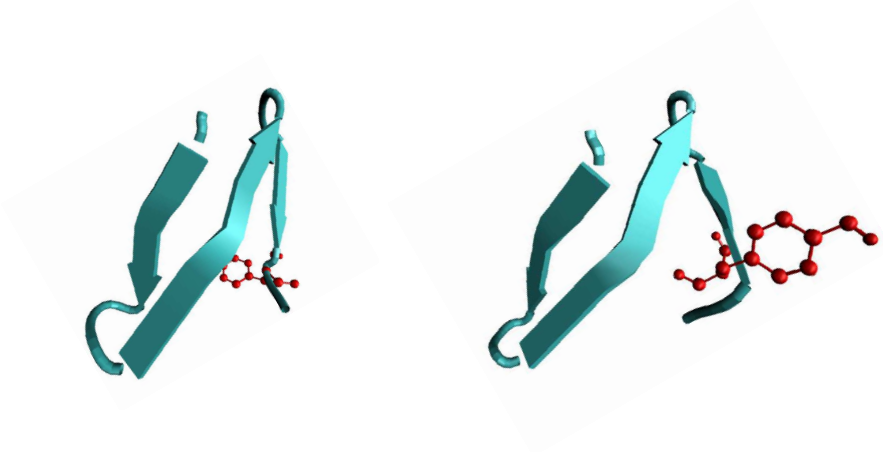


FIG. S5: Two snapshots belonging to the native secondary structure string, despite a completely different orientation of the Tyrosine19 sidechain.

E. Barriers in the entropic region

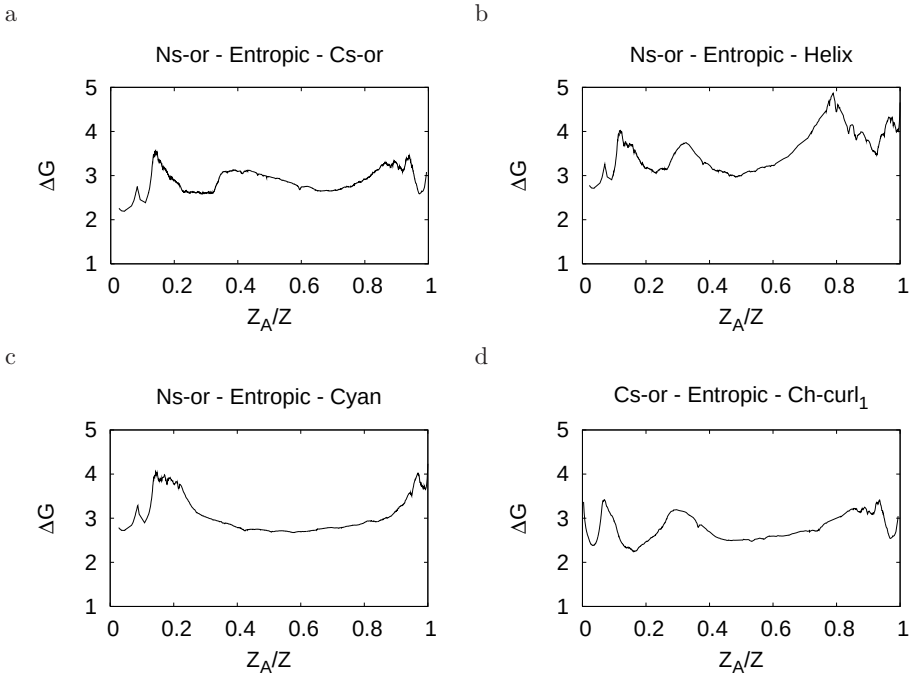


FIG. S6: Reduced pfoldf profiles. Only two enthalpic basins plus the entropic region are used to plot these FEPs. The entropic region, which stretches between the first (second in the case of Ns-or) and the last barrier, reveals barriers (a, b, d) that are otherwise invisible. The pairs of basins were chosen such that very few (or no) direct transitions between them were observed in the simulation except for Ns-or/cyan. Secondary structure-based coarse-graining was used for these profiles.

---

* corresponding authors, tel: +33 390 24 5123 fax: +33 390 24 5124, e-mail: `marci@tammy.harvard.edu,caflisch@bioc.uzh.ch`

† SK and SM have made equal contributions to this study.

[1] S. V. Krivov and M. Karplus, *J. Phys. Chem. B*, **2006**, *110*, 12689–12698.

[2] S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. USA.*, **2004**, *101*, 14766–14770.

Chapter 5

Identification of the protein folding transition state from molecular dynamics trajectories

[*Submitted*]

Identification of the protein folding transition state from molecular dynamics trajectories

S. Muff and A. Caflisch*

Department of Biochemistry,

University of Zurich,

Winterthurerstrasse 190,

CH-8057 Zurich, Switzerland

tel: +41 44 635 55 21,

fax: +41 44 635 68 62,

e-mail: caflisch@bioc.uzh.ch

(Dated: December 9, 2008)

Abstract

The rate of protein folding is governed by the transition state, so that a detailed characterization of its structure is essential for understanding the folding process. *In vitro* experiments have provided a coarse-grained description of the folding transition state ensemble (TSE) of small proteins. Atomistic details could be obtained by molecular dynamics (MD) simulations but it is not straightforward to extract the TSE directly from the MD trajectories, even for small peptides. Here, the structures in the TSE are isolated by the cut-based free-energy profile (cFEP) using the network whose nodes and links are configurations sampled by MD and direct transitions among them, respectively. The cFEP is a barrier-preserving projection that does not require arbitrarily chosen progress variables. First, a simple two-dimensional free-energy surface is used to illustrate the successful determination of the TSE by the cFEP approach, and to explain the difficulty in defining boundary conditions of the Markov state model for an entropically stabilized free-energy minimum. The cFEP is then used to extract the TSE of a β-sheet peptide with a complex free-energy surface containing multiple basins and an entropic region. In contrast, Markov state models with boundary conditions defined by projected variables and conventional histogram-based free energy profiles are not able to identify the TSE of the β-sheet peptide.

Keywords: molecular dynamics, transition state ensemble, p_{fold}, free-energy profile, denatured state ensemble

I. INTRODUCTION

Proteins fold from the heterogeneous set of denatured conformations to the structurally well-defined native state by a complex conformational transition governed by the free-energy surface[1]. In remarkable contrast to the complexity of the folding process, a simple two-state description, i.e., folded and denatured free-energy minima separated by the transition state ensemble (TSE), is often used to describe the experimental measurements on single-domain fast-folding proteins[2]. As a consequence, little information concerning the details of the folding pathways is obtained, although experimental approaches based on mutagenesis have played a key role in providing a description of the residue interactions at the TSE[3]. Also, studies supplementing the kinetic measurements by probes sensitive to structural details[4,5] have shed some light into the folding pathways, particularly when intermediates are present[6]. However, none of the experimental studies can provide a detailed description of the structures that are visited along the folding pathways. In particular, it is difficult to determine the structures of the folding TSE because of their transient character and the many degrees of freedom of the polypeptide chain. In a nutshell, the TSE is elusive and complex.

Several approaches have been proposed to identify putative TSE structures along molecular dynamics (MD) trajectories[7–11]. Moreover, a procedure based on the evaluation of the probability of folding before unfolding (p_{fold}) by additional short simulations[12] (termed p_{fold}^{MD} in the following) has been used for validating putative TSE structures[9,13–15]. Because of its high computational cost the p_{fold}^{MD} approach is used only to validate the TSE and not to extract it from the trajectories. It is important to note that the definition of p_{fold} is the origin of many difficulties when it comes to practical applications, because in contrast to the folded state, which is well-defined by structural criteria, it is all but simple to define the usually very heterogeneous denatured state. An efficient but approximate approach to calculate p_{fold} directly from the original trajectory upon coarse-graining (termed p_{fold}^{N} in the following), i.e., without any additional simulation, does not need the identification of the denatured state[11]. A more accurate way to determine the same quantity is by analytical calculation on the ETN with the pfoldt procedure[16]. Both p_{fold}^{N} and pfoldt require the choice of a commitment time, which is not simple and rather arbitrary if the system is not known in detail.

In this paper we show that the folding TSE structures can be identified accurately by

the cut-based free-energy profile (cFEP)[17] obtained from MD simulations. The cFEP is a barrier-preserving projection onto a progress coordinate that takes into account all routes to leave or enter the free-energy basin chosen as reference. It uses as input the equilibrium transitions network (ETN), i.e., the capacitated graph whose nodes and links represent coarse-grained microstates and transitions, respectively, sampled by MD simulations[16,17]. In particular, the unfolding barrier, which is the barrier to leave the folded state, can be determined exactly by the cFEP. The procedure to isolate the TSE by the cFEP is validated here on a simple and illustrative two-state free-energy surface, as well as on a complex system with 645 degrees of freedom. The latter is a structured peptide (20-residue three-stranded antiparallel β-sheet called Beta3s) for which several folding-unfolding events can be sampled by implicit solvent molecular dynamics simulations[18,19]. In contrast to the cFEP method, the calculation of p_{fold} values within the framework of a Markov state model (p_{fold}^{MSM}) fails to isolate the TSE because it requires the definition of initial (unfolded) and final (folded) regions. These boundary states are not determined adequately by the arbitrary selection of an unfolded state representative or by means of geometric variables like the number of native contacts or the root mean square deviation (rmsd) from the folded structure. Finally, we show that conventional free-energy projections onto apparently appropriate geometrical variables are not useful for determining the TSE of Beta3s.

II. METHODS

Table 1 gives a short description of the procedures that were employed to isolate and validate the TSE, as well as advantages and disadvantages of the approaches (cFEP, p_{fold}^{MD}, p_{fold}^{N}, pfoldt, and p_{fold}^{MSM}) used to determine or validate putative TSE structures. Other approaches to bias simulations towards structures in the TSE[20,21] or to isolate them from unfolding simulations[9] require experimental data (ϕ-values[3]) and are therefore not directly comparable with the procedures used in this paper.

A. Transition state identification from the cut-based free-energy profile (cFEP)

Projected free-energy surfaces are most useful if they preserve the barriers and minima in the order that they are met during folding/unfolding events. Using an analogy between

the system kinetics and equilibrium flow through a network, Krivov and Karplus have introduced the cut-based free-energy profile (cFEP) and a progress coordinate that have most of these properties[17]. The input for the cFEP calculation is the ETN, which is derived from the trajectory of coarse-grained microstates. The progress coordinate is the normalized partition function of the reactant region containing the native node A (Z_A/Z), but other progress coordinates can be used, because the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate[22]. The result is a one-dimensional profile that preserves the barriers between the free-energy basins; given the barriers, the minima can be determined[17]. The method was applied to the β-hairpin of protein G[17] and Beta3s[16].

During the procedure of cFEP calculation, all nodes of the system are assigned a value of the progress variable. Here, the mean first passage time (mfpt) to the folded node was used as progress variable. The latter can be calculated analytically for every node if the ETN fulfils the Markov property. The occurrence of the nodes in the profile is sorted in ascending order of mfpt from the native state (i.e., from the first node along Z_A/Z). Therefore, every node i can be localized along the cFEP according to its progress variable mfpt_i, because the Z_A/Z coordinate that corresponds to node i can be calculated as $Z_A/Z(i) = \sum_{\mathrm{mfpt}(j)\leq\mathrm{mfpt}(i)} Z_j/Z$, where Z_j is the partition function of node j, and Z is the total partition function. Since the cFEP takes into account all routes present in the ETN to and from the initial state[16,17] without any prejudgment as to the geometric coordinates or pathways involved, the TSE is situated on top of the unfolding barrier. In this way, nodes corresponding to the top of the first barrier to exit the native state can automatically be identified as TSE structures. All cFEPs in this work were calculated using the program WORDOM[23].

B. Evaluation of $p_{fold}(\tau_{commit})$ with additional simulations (p_{fold}^{MD})

If snapshots saved along a trajectory are grouped into structurally homogeneous nodes during the coarse-graining procedure, nodes belonging to the TSE have equal probability to fold and to unfold, i.e., $p_{fold} \approx 0.5$, whereas folded and unfolded regions correspond to $p_{fold} \approx 1$ and $p_{fold} \approx 0$, respectively. In order to validate the application of cFEPs to identify the transition, folded, and unfolded ensembles, a large number of MD trajectories from various structures with varying initial distribution of velocities can be started and the fraction of those that fold within a commitment time τ_{commit}[12,13,24] corresponds to the

respective p_{fold}^{MD}. τ_{commit} has to be chosen much longer than the shortest time-scales of conformational fluctuations and much shorter than the average folding time[13].

The p_{fold}^{MD} calculations are computationally very expensive, since the error for a structure scales with $\sqrt{n}$, where n is the number of trajectories started from it. Therefore, the realization of many short trajectories from individual structures cannot be applied directly to identify the TSE from molecular dynamics trajectories, and is used in this work only to validate putative transition state structures as isolated by other approaches.

C. $p_{fold}(\tau_{commit})$ calculation directly from the trajectories (p_{fold}^{N}) or the ETN (pfoldt)

The calculation of p_{fold}^{MD} is computationally very expensive and is feasible only for a small subset of nodes. In a previous work, a method was proposed for estimating folding probabilities for *all* structures visited in an equilibrium folding-unfolding trajectory[11]. The calculation does not require any additional simulations because the original MD trajectory is used to directly estimate the folding probabilities. The τ_{commit}-segment of the MD trajectory following each snapshot is analyzed to check if the folding condition is met (i.e, that the folded node, which usually is the most populated one, is visited). For each node, the ratio between the snapshots which lead to folding and the total number of snapshots in the node is defined as the node-p_{fold} (p_{fold}^{N}). This value is an approximation of the $p_{fold}(\tau_{commit})$ of any single structure in the node which is valid if the node consists of structurally and kinetically similar conformations. The error in p_{fold}^{N} scales with $\sqrt{W}$, where W is the number of structures in the node.

The analytical calculation of $p_{fold}(\tau_{commit})$ on the ETN of a Markov state model, termed pfoldt, is more accurate than p_{fold}^{N} because it uses the full connectivity of the ETN, thereby reducing the statistical error. The corresponding equation system was introduced previously[16] and the results were used as a progress variable for cFEP calculations (pfoldt procedure).

D. p_{fold} calculation with a Markov state model (p_{fold}^{MSM})

Within the framework where the ETN corresponds to a Markov state model, folding probabilities can be calculated directly, if the two regions U (unfolded) and F (folded) are known. Given these regions, the folding probability of a node i within the Markov state model is found as the solution of the equation system $p_i^{\mathrm{MSM}} = \sum_j p_{ji} p_j^{\mathrm{MSM}}$ with boundary conditions $p_{\kappa\in U}^{\mathrm{MSM}} = 0$ and $p_{\kappa\in F}^{\mathrm{MSM}} = 1$ Refs.[10,25]. The equation system can be solved efficiently by iterative multiplication of the vector p_j^{MSM} by the matrix p_{ji}. According to the cFEP, the folded and unfolded states are defined as all nodes on the left and right of the folding barrier, respectively. However, use of this definition to determine U and F would be tautological, because if the cFEP is known, there is no need for the Markov state model approach, and it is then trivial that the nodes on the barrier have no other choice than attaining $p_{fold} \approx 0.5$. Therefore, the Markov state model approach is applied in this work without the input of the knowledge from the cFEP in order to objectively compare p_{fold}^{MSM} with the other methods.

III. TWO-STATE SYSTEM WITH ENTROPIC FREE-ENERGY MINIMUM: AN ILLUSTRATIVE MODEL

A simple two-dimensional, radially symmetric potential energy surface illustrates the correct TSE isolation by the cFEP, and the dependency of the p_{fold}^{MSM} on the definition of initial and final regions. The corresponding free-energy surface has only two minima: An enthalpic, funnel-like "folded" state and a purely entropic "denatured" state[26] (Fig. 1A). The discretization of the simple potential yields an ETN[26] (Fig. 1B), which is similar to, but much simpler than, what is usually obtained from the coarse-graining procedure of simulations of peptides (and proteins). There is a free-energy barrier at r = 0 that separates the enthalpic (r < 0) from the entropic basin (r > 0), as indicated by the green line in Figs. 1A and B.

In a first step, it was verified that the cFEP is able to identify the TSE. The TSE is correctly grouped around the barrier in the cFEP and only the nine nodes in the neighborhood of the minimal cut (at r $\lessapprox$ 0) lie between the first and the last green circle in the cFEP (Fig. 1C). In a next step, p_{fold}^{MSM} was calculated between the most populated node (black, $p_{fold}^{\mathrm{MSM}} = 1$) and an arbitrary representative of the entropic state (red, $p_{fold}^{\mathrm{MSM}} = 0$). The

strong dependency of the $p_{fold}^{\mathrm{MSM}} \approx 0.5$ region on the choice of the unfolded representative is remarkable (Fig. 1D-F). Furthermore, none of the choices is able to fully reveal the correct TSE, although the system is very simple. This result illustrates the difficulty of selecting a representative structure, which is in practice almost impossible in the case of an entropically stabilized state.

IV. APPLICATION TO BETA3S

In the previous example the complete knowledge about the free-energy surface is available and it is very simple to correctly determine the folded and unfolded regions and therefore also the TSE. However, this is an oversimplified and unrealistic case, and the following application to the structured peptide Beta3s illustrates the advantage of the cFEP approach for the analysis of complex systems.

A. MD simulations

Beta3s is a designed 20-residue sequence whose solution conformation has been investigated by NMR spectroscopy[27]. The NMR data indicates that Beta3s in water forms a monomeric (up to more than 1mM concentration) triple-stranded antiparallel β-sheet, in equilibrium with the denatured state[27]. We have previously shown that in implicit solvent[28] molecular dynamics simulations Beta3s folds reversibly to the NMR solution conformation, irrespective of the starting structure[18]. Recently, analysis of a 20-μs equilibrium MD simulation close to the melting temperature at 330 K revealed a very heterogeneous denatured state with a large entropic region and multiple enthalpic traps[16,19,29]. The same 20-μs of MD sampling was used here, and there are a total of 10^6 snapshots because coordinates were saved every 20 ps. The simulations were performed with the program CHARMM[30]. Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[30]). A mean field approximation based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute[28]. The two surface tension-like parameters of the solvation model were optimized without using Beta3s. The same force field and implicit solvent model have been used in molecular dynamics simulations of the early steps of ordered aggregation[31], and folding of

structured peptides[18,28,32], as well as small proteins of about 60 residues[33].

B. Coarse-graining

The leader algorithm[34] is used for coarse-graining the snapshots according to the all-atom rmsd and a cutoff value of 2.5 Å[16]. The current snapshot is grouped to the last (in time) visited node whose central snapshot has an rmsd from the current snapshot lower than the cutoff. This version of the leader algorithm accounts not only for structural, but also for kinetic similarity, because recently visited snapshots are more likely to be kinetically close than those that were visited with a large temporal delay. The importance of using transitions rather than only structures to assign states has been recently investigated for a small helical peptide[35].

Note that nodes in the ETN with only one or two neighbors (i.e., one incoming and/or one outgoing neighbor) were grouped to their outgoing neighbor. This regrouping is justified because the future of such nodes within a trajectory is completely determined, i.e., no information is erased through their regrouping. Upon rmsd coarse-graining and regrouping 34'671 nodes and 151'819 links were visited. These nodes are the states of the Markov state model (i.e., the ETN), and the lagtime was set to Δt=20 ps. Figure 2 contains a comparison of folding dynamics from the MD simulations and from the corresponding Markov state model at 330 K. Essentially the same kinetics can be extracted, indicating that the Markov assumption holds and non-Markovian noise is negligible.

C. Correct identification of the TSE by the cFEP method

The native basin is bounded by the first local maximum in the unfolding cFEP of Beta3s, which is the cFEP with the native node as reference (Fig. 3). To show that the TSE is situated on top of the unfolding barrier, the folding probability p_{fold}^{MD} with a τ_{commit} value of 5 ns was evaluated on 34 nodes by running additional MD simulations (see Methods). These nodes were selected equally spaced along the Z_A/Z, except for a higher density in a region bracketing the unfolding barrier on the cFEP (results with mfpt as progress coordinate are shown in Supp. Mat. Fig. S1). Ten structures were chosen randomly from every node, and 20 simulations of 10 ns each with different initial velocities were started from each structure, i.e.,

a total of 200 simulations was accumulated per node. Note that for this validation it is more informative to monitor, for individual snapshots, the monotonously growing behavior of p_{fold}^{MD} as a function of τ_{commit} rather than selecting a single value of τ_{commit}[15] (Fig. 4). Notably, the top of the first free-energy barrier in the cFEP corresponds to the $p_{fold}^{\mathrm{MD}} \approx 0.5$ region, i.e., to the folding/unfolding TSE of Beta3s (Fig. 3A). The accuracy of the TSE identification on the cFEP is striking. Nodes before the first barrier belong to the native basin and have $p_{fold}^{\mathrm{MD}} \approx 1$, while nodes after the barrier have $p_{fold}^{\mathrm{MD}} \approx 0$. Moreover, the distributions of p_{fold}^{MD} values over the ten structures in each node are peaked around the respective average p_{fold}^{MD} value, even for TSE nodes (Supp. Mat. Fig. S2). These results show that the cFEP approach is able to correctly identify not only free-energy basins and barriers of complex systems, but also the TSE to exit or enter the region of interest (usually the folded state). The very good correlation between the increasing values of Z_A/Z along the cFEP and p_{fold}^{MD} can be explained because both mfpt and p_{fold}^{MD} describe the kinetic distance from a state. Note that this correlation is not due to a tautology because the cFEP is calculated using the ETN as input, whereas p_{fold}^{MD} is extracted from additional MD simulations. Moreover, the correlation is robust with respect to the choice of the progress variable of the cFEP procedure (see Supp. Mat. Fig. S3).

Importantly, no additional parameter is needed to calculate the cFEP, which only requires the selection of the native node (or the representative node of any other free-energy basin). Note that the perfect match between the unfolding barrier on the cFEP and the sharp decay of the p_{fold}^{MD} values justifies a posteriori the choice of a commitment time of 5 ns used to calculate the latter. Essentially identical results are obtained with a commitment time of 10 ns (Table II and Supp. Mat. Fig. S4).

D. Approximation by p_{fold}^{N} and pfoldt

As mentioned in the Methods section, the calculation of p_{fold}^{MD} is computationally very expensive and is feasible only for a small subset of nodes. On the other hand, the calculation of p_{fold}^{N} does not require any additional simulations[11]. The p_{fold}^{N} values, which were also calculated with $\tau_{commit} = 5$ ns, are close to the p_{fold}^{MD} for most of the 34 nodes used to calculate the latter (Fig. 3B, and Table II). The error is due to low statistics harvested for the p_{fold}^{N} estimation, which is limited by the number of visits to the node within the trajectory. This

problem is particularly severe for TSE nodes, which suffer from low sampling, whereas highly populated conformations can be classified more reliably (Fig. 5, left).

The results improve dramatically if p^{N}_{fold} is replaced by its equivalent calculated on the ETN, i.e., pfoldt (Fig. 3C), and a very sharp decay of folding probabilities at the cFEP unfolding barrier can be observed. Pfoldt is significantly more accurate than p^{N}_{fold} because as mentioned above the ETN contains all connectivity and pathway information between regions, which is not present if only isolated sampling events from individual nodes during the simulation are considered (as for p^{N}_{fold}). Results for all nodes populated by a significant number of snapshots above a certain cutoff are given in Fig. 5, middle. Again, the comparison of p^{N}_{fold} (left panel) with pfoldt (middle panel) confirms the higher accuracy of the latter. Note that pfoldt, like p^{N}_{fold} and p^{MD}_{fold}, relies on the correct choice of τ_{commit}, a parameter that is usually not simple to determine.

E. Failure of TSE identification by a Markov state model with boundaries defined according to structural criteria

The p_{fold} calculation in the framework of a Markov state model (p^{MSM}_{fold}) involves the choice of representative regions for the folded and unfolded state with $p_{fold} = 1$ and $p_{fold} = 0$, respectively, as boundary conditions[10,25,36]. Like in many previously published applications, the regions U and F were determined according to a simple structural criterion based on the number of native contacts Q, which is a commonly used geometrical variable. A node was assigned to the initial and final region, if the structures in that node had on average less than 5 or more than 19 of the 26 native contacts[18] formed, respectively (Supp. Mat. Fig. S5). With this definition of boundary conditions, U and F consist of 31% and 27% of the total number of snapshots, respectively.

Calculation of p^{MSM}_{fold} values from the equation system reveals that several nodes after the unfolding barrier have a $p^{\mathrm{MSM}}_{fold} > 0.5$ (Fig. 3D). Fig. 6 contains as a supplement to the cFEP the location of all putative TSE-nodes as isolated with p^{MD}_{fold}, p^{N}_{fold}, pfoldt, and p^{MSM}_{fold}, i.e., those nodes with $0.45 < p_{fold} < 0.55$ and at least 20 snapshots for statistical significance. While all three τ_{commit}-based methods approximate the TSE quite well (Fig. 6A-C), most structures identified by p^{MSM}_{fold} are far away from the unfolding barrier in the cFEP (Fig. 6D). The incorrect determination of the TSE by p^{MSM}_{fold} is also shown for the regions U and F

defined by all-atom rmsd > 5.5 Å (weight of 48%) and all-atom rmsd < 2.5 Å (weight of 25%), respectively (Fig. S6 of Supp. Mat.).

The isolation of the correct TSE by p_{fold}^{MSM} can only be achieved if the selected regions are "true" representatives of the folded and unfolded states, i.e., if each time the polypeptide folds or unfolds (and only then), the folded or unfolded region is visited, respectively. It is important to emphasize that, except for a two-state system with well-defined native and non-native basins, the choice of such representative ensembles is very difficult and mostly impossible by geometrical criteria. This problem originates from the usually very heterogeneous character of the denatured state with multiple basins and/or an entropic region[16,19,26]. While the representation of the folded state by a single node may be legitimate if the basin is enthalpic, the denatured state cannot be represented by a single node. For instance, for each choice of the unfolded representative disparate $p_{fold}^{\mathrm{MSM}} \approx 0.5$ regions are obtained for Beta3s (Supp. Mat. Fig. S7).

F. Failure of TSE identification from free-energy projections onto geometric variables

In a previous work, the number of native contacts in the N-terminal hairpin (Q_N) and C-terminal hairpin (Q_C) of Beta3s were used as progress variables to investigate thermodynamics and folding pathways sampled by MD simulations close to the melting temperature[18]. Note that these variables are the most "natural" among the geometric coordinates, considering that a three-stranded antiparallel β-sheet has an inherent symmetry and consists of two β-hairpins sharing the central β-strand. The histogram-based projection of the free energy onto the (Q_N, Q_C)-plane showed two barriers separating the folded from the denatured state at ($Q_N = 4/11$, $Q_C = 9/11$) and ($Q_N = 10/11$, $Q_C = 3/11$), with the former lower by about 0.5 kcal/mol than the latter as shown in Supp. Mat. Fig. S8. To calculate p_{fold}^{MD}, multiple short MD runs were started from 10 structures with ($Q_N = 4/11$, $Q_C = 9/11$) and 10 structures with ($Q_N = 10/11$, $Q_C = 3/11$). The value of p_{fold}^{MD} was equal (or very close) to 1 or 0 for 19 of the 20 putative TSE structures (data not shown). This failure is not surprising considering the sharp decay of p_{fold}^{MD} at the cFEP barrier (Fig.3A), which suggests that the correct identification of the TSE is very sensitive and not possible at all if the choice of the progress variable(s) results in projections that do not preserve the barrier(s). Therefore,

free-energy projections onto geometric variables are in general not appropriate to determine the folding TSE.

V. CONCLUSIONS

The accurate determination of the TSE is essential for understanding the protein folding reaction. This paper deals with the automatic extraction of folding TSE structures for a simple two-dimensional energy surface and from MD simulations of a structured peptide. The cFEP, a barrier-preserving projection able to fully quantify the kinetic and thermodynamic properties of a system at equilibrium[17], is shown to successfully determine TSE conformations at the top of the transition region to enter or leave a free-energy basin. On the other hand, free-energy projections onto geometric coordinates like the fraction of native contacts or the rmsd from the native structure are shown to fail (for the structured peptide) as most of the conformations at the maxima of the projected surface do not belong to the TSE. This failure is a consequence of the sharpness of the folding transition barrier and the fact that such projections do not preserve the location of the barriers. The TSE determination has been attempted previously only for minimally frustrated systems[14,37,38], or for reactions involving a small and well-defined region of a protein[39]. For such reactions, an automatic procedure can identify reaction coordinates from an initial guess of several thousands physical variables, but requires the evaluation of commitment probabilities by additional simulations[40], which is computationally prohibitive for a large set of structures.

In contrast to the automatic and parameter-free TSE determination by the cFEP, conventional p_{fold}-based methods involve the choice of a commitment time, or the arbitrary selection of representative regions for the native and the denatured state. The TSE isolation from the original MD trajectory (p^{N}_{fold})[11] or by analytical calculation on the ETN (pfoldt)[16] are very efficient and do not require any additional simulations, but the results depend on the choice of the commitment time. Moreover, p^{N}_{fold} values can be biased if insufficient amount of statistics are harvested, especially at the the transition region, which is naturally sampled less than the free-energy minima.

More problematic is the p_{fold} calculation with a Markov state model (p^{MSM}_{fold}), because for a complex free-energy surface it is not possible to define the boundary conditions (i.e., $p^{\mathrm{MSM}}_{fold} = 0$ and 1) by simple structural criteria. This implies that most choices of such

boundary regions lead to wrong p_{fold}^{MSM} results and thus to a flawed or incomplete isolation of the TSE. It is important to note that the same coarse-graining of the structures and ETN are employed in the p_{fold}^{MSM} calculation and the cFEP approach, but only the latter does not require that the denatured state is defined a priori.

The difficulty related to the calculation of p_{fold} lies in nuances of its definition: p_{fold} is the probability to fold before unfolding[12]. While p_{fold} calculated using a commitment time approximates this definition, p_{fold}^{MSM} between regions F and U is the probability to visit region F before U, which corresponds to the original definition of p_{fold} only if the trajectory visits F and U each time it folds and unfolds, respectively, but not in between. Therefore, it is likely that p_{fold}^{MSM} calculations will be valid only in very special cases, e.g., in a two-state system with two enthalpic basins, where (simple) geometrical criteria are sufficient to separate the states. In contrast, the cFEP is able to isolate the TSE from a complex free-energy surface and does not necessarily require (long) equilibrium folding-unfolding simulations, as recently shown for an ETN obtained by short segments of replica exchange MD trajectories[41].

Acknowledgments

We thank Dr. R. Pellarin and P. Schütz for critical reading of the manuscript. We thank Dr. A. Cavalli for performing some of the simulations for p_{fold}^{MD} evaluation. The molecular dynamics simulations were performed on the Etna and Matterhorn computer clusters at the University of Zurich. This work was supported by a Swiss National Science Foundation grant to A.C.

---

* corresponding author, tel: +41 44 635 55 21, fax: +41 44 635 68 62, e-mail: `caflisch@bioc.uzh.ch`

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).

[2] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10248 (1991).

[3] A. Matouschek, J. T. Kellis, Jr., L. Serrano, and A. R. Fersht, Nature **340**, 122 (1989).

[4] C. M. Dobson, A. Šali, and M. Karplus, Angew. Chem. Int. Ed. **37**, 869 (1998).

[5] J. A. Ihalainen *et al.*, Proc. Natl. Acad. Sci. USA. **104**, 5383 (2007).

[6] S. E. Radford, C. M. Dobson, and P. A. Evans, Nature **358**, 302 (1992).

[7] A. Li and V. Daggett, Proc. Natl. Acad. Sci. USA. **91**, 10430 (1994).

[8] L. Li and E. I. Shakhnovich, Proc. Natl. Acad. Sci. USA. **98**, 13014 (2001).

[9] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. USA. **99**, 6719 (2002).

[10] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).

[11] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch, J. Chem. Phys. **122**, 184901 (2005).

[12] R. Du *et al.*, J. Chem. Phys. **108**, 334 (1998).

[13] I. Hubner, J. Shimada, and E. Shakhnovich, J. Mol. Biol. **336**, 745 (2004).

[14] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA. **103**, 586 (2006).

[15] G. Settanni and A. Fersht, Biophys. J. **94**, 4444 (2008).

[16] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).

[17] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).

[18] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. USA. **97**, 10780 (2000).

[19] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).

[20] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, Nature **409**, 641 (2001).

[21] G. Settanni, J. Gsponer, and A. Caflisch, Biophys. J. **86**, 1691 (2004).

[22] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. USA. **105**, 13841 (2008).

[23] M. Seeber *et al.*, Bioinformatics **23**, 2625 (2007).

[24] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).

[25] W. Swope, J. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).

[26] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. USA. **101**, 14766 (2004).

[27] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez, Protein Science **8**, 854 (1999).

[28] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Structure, Function, and Bioinformatics **46**, 24 (2002).

[29] S. Muff and A. Caflisch, Proteins: Structure, Function, and Bioinformatics **70**, 1185 (2008).

[30] B. R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983).

[31] J. Gsponer, U. Haberthür, and A. Caflisch, Proc. Natl. Acad. Sci. USA. **100**, 5154 (2003).

[32] J. A. Ihalainen *et al.*, Proc. Natl. Acad. Sci. USA. **105**, 9588 (2008).

[33] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).

[34] J. Hartigan, Wiley, New York (1975).

[35] N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).

[36] F. Noe and S. Fischer, Curr. Opin. Struct. Biol. **18**, 154 (2008).

[37] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. USA. **97**, 5877 (2000).

[38] R. Best and G. Hummer, Proc. Natl. Acad. Sci. USA. **102**, 6732 (2005).

[39] J. Hu, A. Ma, and A. R. Dinner, Proc. Natl. Acad. Sci. USA. **105**, 4615 (2008).

[40] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).

[41] S. Muff and A. Caflisch, submitted to JPC B (2008).

Procedures used to validate or identify the folding TSE.

Procedure	Data used	Advantages	Disadvantages[a]	Reference
cFEP	ETN	fast exact on ETN	requires coarse-graining	[16,17]
p_{fold}^{MD}	additional MD runs	exact, used for validation no coarse-graining	computationally expensive requires τ_{commit}	[12]
p_{fold}^{N}	original trajectory	fast	requires τ_{commit} strong dependency on sampling requires coarse-graining	[11]
pfoldt	ETN	fast exact on ETN	requires τ_{commit} requires coarse-graining	[16]
p_{fold}^{MSM}	ETN	fast	$p_{fold}^{\mathrm{MSM}} = 0.5 \not\Rightarrow$ folding-TSE requires unfolded state definition requires coarse-graining	[10,25]

TABLE I: Abbreviations: cFEP, cut-based free-energy profile; ETN, equilibrium transitions network; MSM, Markov state model; τ_{commit}, commitment time. [a]The cFEP, p_{fold}^{N}, pfoldt, and p_{fold}^{MSM} methods rely on sufficient sampling and a meaningful coarse-graining of the trajectories. Note that the p_{fold}^{N} procedure has a stronger dependency on sampling than cFEP, pfoldt, and p_{fold}^{MSM}. The latter procedures use the ETN, which is much more informative than the original trajectory itself because the ETN represents the complete connectivity information of all states

Z_A/Z	p_{fold}^{MD}(5 ns)	p_{fold}^{MD}(10 ns)	p_{fold}^{N}(5 ns)	p_{fold}^{N}(10 ns)	pfoldt(5 ns)	pfoldt(10 ns)	p_{fold}^{MSM}	Node population
0.2	0.995	0.995	0.673	0.878	0.997	0.998	1.000	539
0.25	0.995	1.000	0.966	0.979	0.997	0.998	1.000	726
0.3	0.915	0.945	0.923	0.962	0.984	0.990	0.504	26
0.3500	0.675	0.815	0.792	0.917	0.971	0.981	0.430	24
0.3525	0.845	0.900	0.895	1.000	0.964	0.975	0.594	19
0.3550	0.645	0.750	0.656	0.656	0.951	0.967	0.359	32
0.3575	0.735	0.780	0.500	0.500	0.910	0.942	0.576	34
0.3600	0.905	0.925	0.583	0.833	0.872	0.916	0.459	12
0.3625	0.630	0.700	0.600	0.600	0.806	0.874	0.417	16
0.3650	0.630	0.705	0.333	0.733	0.740	0.834	0.728	15
0.3675	0.665	0.730	0.333	0.611	0.704	0.805	0.425	18
0.3700	0.345	0.540	0.364	0.364	0.588	0.752	0.560	11
0.3725	0.260	0.405	0.000	0.000	0.481	0.687	0.636	27
0.3750	0.215	0.310	0.000	0.000	0.329	0.572	0.325	15
0.3775	0.235	0.360	0.800	0.867	0.390	0.603	0.475	15
0.3800	0.220	0.350	0.116	0.116	0.239	0.524	0.618	147
0.3825	0.250	0.360	0.909	0.909	0.248	0.542	0.354	22
0.3850	0.055	0.145	0.000	0.688	0.003	0.570	0.723	16
0.3875	0.110	0.235	0.105	0.105	0.146	0.459	0.556	19
0.3900	0.050	0.140	0.000	0.000	0.225	0.459	0.652	177
0.3925	0.035	0.095	0.000	0.000	0.003	0.463	0.555	10
0.3950	0.110	0.155	0.000	0.000	0.052	0.361	0.161	10
0.4000	0.020	0.055	0.000	0.015	0.000	0.460	0.481	66
0.45	0.015	0.030	0.029	0.126	0.000	0.271	0.446	8584
0.5	0.045	0.135	0.044	0.134	0.000	0.251	0.459	14918
0.55	0.040	0.135	0.281	0.509	0.000	0.000	0.527	377
0.6	0.010	0.050	0.000	0.000	0.000	0.000	0.628	13
0.65	0.000	0.030	0.000	0.000	0.000	0.000	0.000	16
0.7	0.010	0.045	0.000	0.000	0.000	0.000	0.000	36
0.75	0.005	0.025	0.000	0.000	0.000	0.000	0.000	14
0.8	0.005	0.035	0.000	0.000	0.000	0.000	0.000	25
0.85	0.000	0.020	0.000	0.000	0.000	0.000	0.000	17
0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10
0.95	0.005	0.010	0.000	0.000	0.000	0.000	0.000	14

TABLE II: p_{fold} values of nodes used for the calculation of p_{fold}^{MD}. In the region of the first (i.e., unfolding) barrier of the cFEP, $0.35 \leq Z_A/Z \leq 0.4$, the correlation between p_{fold}^{N} and p_{fold}^{MD} is 0.70, and the correlation between pfoldt and p_{fold}^{MD} is 0.95. Within the same range, there is no correlation between p_{fold}^{MSM} and p_{fold}^{MD} (correlation coefficient of 0.01). Similar correlation coefficients are obtained for $\tau_{commit} = 5$ ns and 10 ns.

Z_A/Z	p_{fold}^{MD}(5 ns)	p_{fold}^{MD}(10 ns)	p_{fold}^{N}(5 ns)	p_{fold}^{N}(10 ns)	pfoldt(5 ns)	pfoldt(10 ns)	p_{fold}^{MSM}	Node population
0.2	0.995	0.995	0.673	0.878	0.997	0.998	1.000	539
0.25	0.995	1.000	0.966	0.979	0.997	0.998	1.000	726
0.3	0.915	0.945	0.923	0.962	0.984	0.990	0.504	26
0.3500	0.675	0.815	0.792	0.917	0.971	0.981	0.430	24
0.3525	0.845	0.900	0.895	1.000	0.964	0.975	0.594	19
0.3550	0.645	0.750	0.656	0.656	0.951	0.967	0.359	32
0.3575	0.735	0.780	0.500	0.500	0.910	0.942	0.576	34
0.3600	0.905	0.925	0.583	0.833	0.872	0.916	0.459	12
0.3625	0.630	0.700	0.600	0.600	0.806	0.874	0.417	16
0.3650	0.630	0.705	0.333	0.733	0.740	0.834	0.728	15
0.3675	0.665	0.730	0.333	0.611	0.704	0.805	0.425	18
0.3700	0.345	0.540	0.364	0.364	0.588	0.752	0.560	11
0.3725	0.260	0.405	0.000	0.000	0.481	0.687	0.636	27
0.3750	0.215	0.310	0.000	0.000	0.329	0.572	0.325	15
0.3775	0.235	0.360	0.800	0.867	0.390	0.603	0.475	15
0.3800	0.220	0.350	0.116	0.116	0.239	0.524	0.618	147
0.3825	0.250	0.360	0.909	0.909	0.248	0.542	0.354	22
0.3850	0.055	0.145	0.000	0.688	0.003	0.570	0.723	16
0.3875	0.110	0.235	0.105	0.105	0.146	0.459	0.556	19
0.3900	0.050	0.140	0.000	0.000	0.225	0.459	0.652	177
0.3925	0.035	0.095	0.000	0.000	0.003	0.463	0.555	10
0.3950	0.110	0.155	0.000	0.000	0.052	0.361	0.161	10
0.4000	0.020	0.055	0.000	0.015	0.000	0.460	0.481	66
0.45	0.015	0.030	0.029	0.126	0.000	0.271	0.446	8584
0.5	0.045	0.135	0.044	0.134	0.000	0.251	0.459	14918
0.55	0.040	0.135	0.281	0.509	0.000	0.000	0.527	377
0.6	0.010	0.050	0.000	0.000	0.000	0.000	0.628	13
0.65	0.000	0.030	0.000	0.000	0.000	0.000	0.000	16
0.7	0.010	0.045	0.000	0.000	0.000	0.000	0.000	36
0.75	0.005	0.025	0.000	0.000	0.000	0.000	0.000	14
0.8	0.005	0.035	0.000	0.000	0.000	0.000	0.000	25
0.85	0.000	0.020	0.000	0.000	0.000	0.000	0.000	17
0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10
0.95	0.005	0.010	0.000	0.000	0.000	0.000	0.000	14

TABLE II: p_{fold} values of nodes used for the calculation of p_{fold}^{MD}. In the region of the first (i.e., unfolding) barrier of the cFEP, $0.35 \leq Z_A/Z \leq 0.4$, the correlation between p_{fold}^{N} and p_{fold}^{MD} is 0.70, and the correlation between pfoldt and p_{fold}^{MD} is 0.95. Within the same range, there is no correlation between p_{fold}^{MSM} and p_{fold}^{MD} (correlation coefficient of 0.01). Similar correlation coefficients are obtained for $\tau_{commit} = 5$ ns and 10 ns.

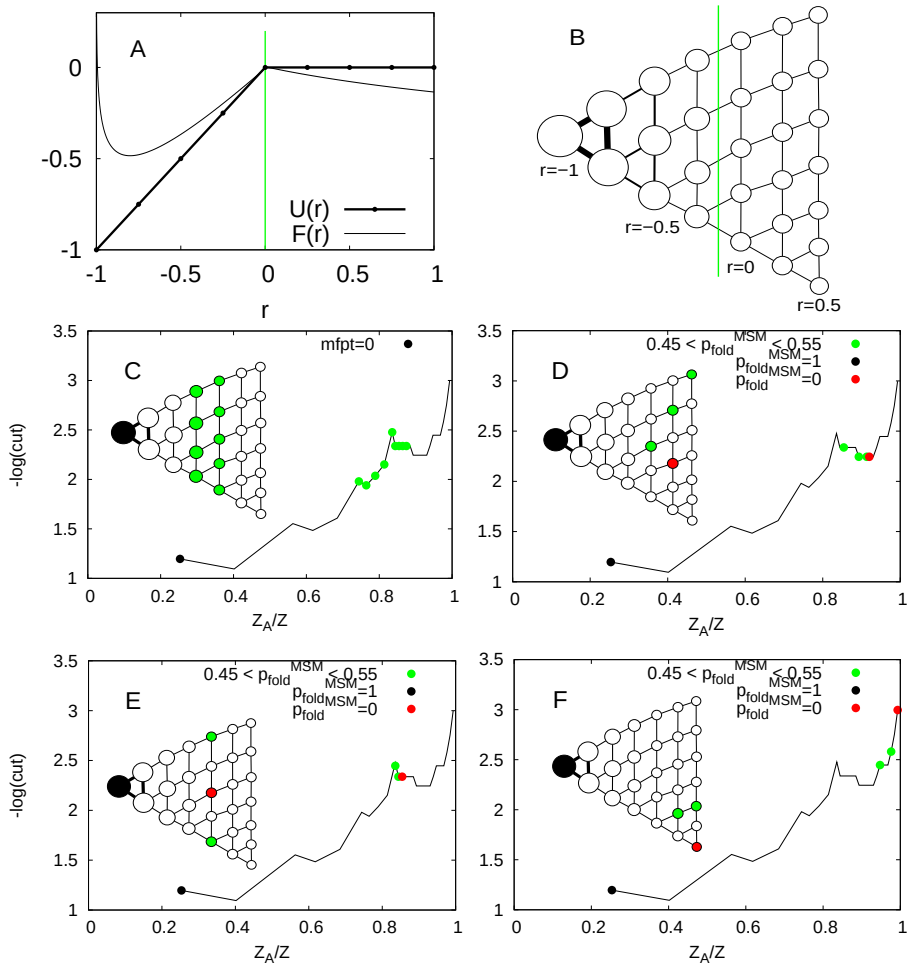


FIG. 1: Simple two-state system with entropically stabilized "unfolded" state[26]. (A) Two-dimensional radially symmetric potential energy surface U(r) and free-energy surface F(r). There are two free-energy minima, one relatively deep representing the enthalpic ("native") state ($r < 0$), and another shallow representing the entropic "unfolded") basin ($r > 0$). The green line indicates the transition state. (B) Discretized ETN of the potential-energy landscape. The size of the nodes and links in the model network is proportional to the partition function of the nodes and transitions. The green line represents the minimal cut through the free-energy barrier, that is, the transition state. (C) The TSE is correctly identified by the cFEP approach, i.e., it consists of the nine nodes on the left and right of the minimal cut in the ETN. (D-F) The solution of the p_{fold}^{MSM} calculations and identification of $p_{fold}^{\mathrm{MSM}} \approx 0.5$ regions is strongly dependent on the node chosen as representative of the entropic region. By none of the three choices it is possible to isolate the complete TSE region correctly. This illustrative model shows that the arbitrary selection of representative nodes in the entropic state is not valid in general and that cFEPs are not affected by this problem because they require only the definition of the native node.

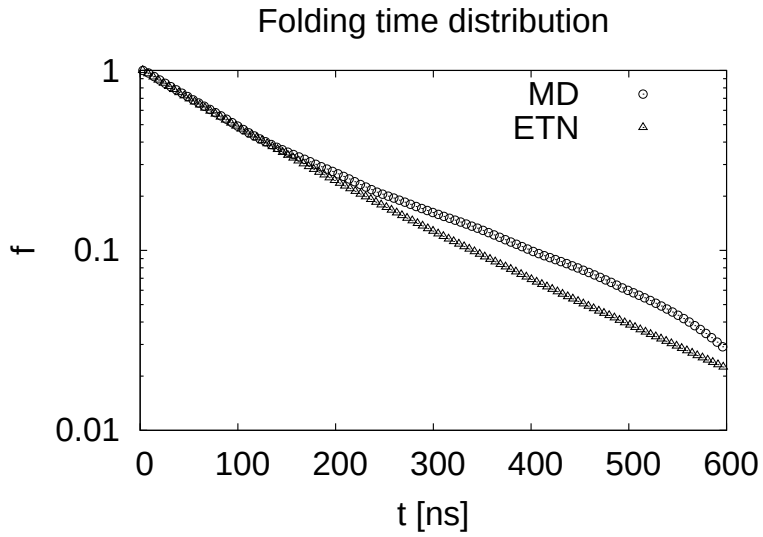


FIG. 2: Cumulative folding time distribution $f(t) = \int_t^\infty p(\tau)d\tau$ as extracted directly from the 20 μs equilibrium simulation of Beta3s (circles) and from the corresponding ETN, which is treated as a Markov state model (triangles). The folding dynamics from the non-native ensemble can be reproduced by the model, which is a strong indication that the Markov assumption is justified for the lagtime of 20 ps used here.

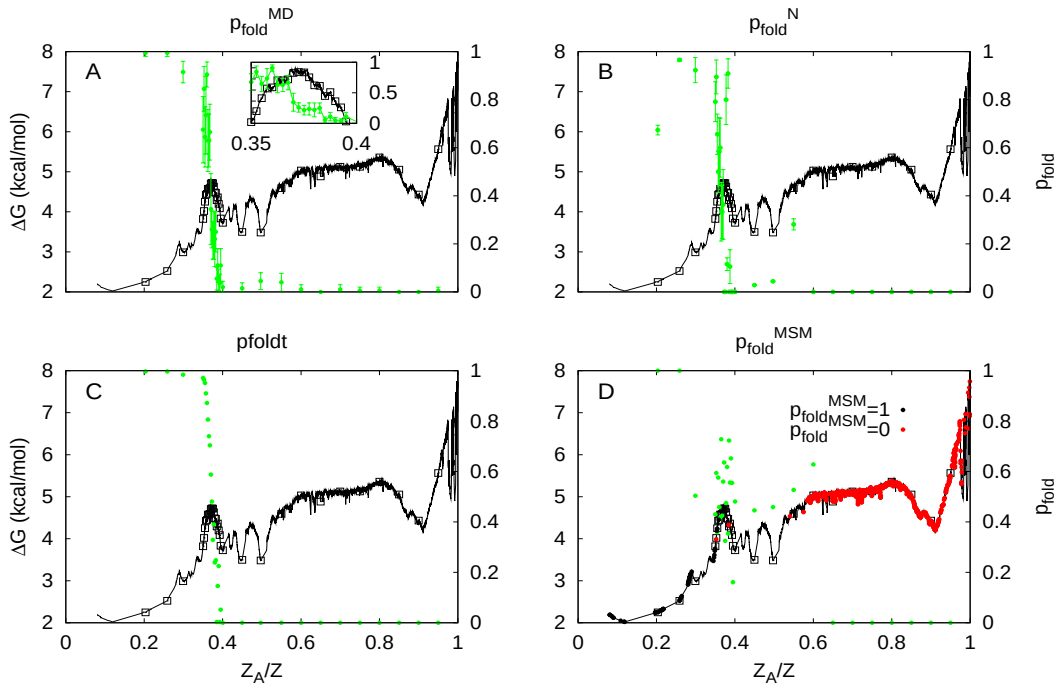


FIG. 3: The TSE can be identified by the cFEP (solid line). The cFEP is shown together with the 34 nodes that were selected for p_{fold}^{MD} calculations (black squares). They are equally spaced along the progress coordinate Z_A/Z, with a distance of 0.05 units except for the first barrier, i.e., 0.35< Z_A/Z <0.4, where the spacing is 0.0025 units to obtain a higher resolution. Values of p_{fold} (green circles) refer to the y-axis on the right. (A) p_{fold}^{MD}. The error bars represent the standard deviation among the ten structures within a node. The inset illustrates the sharp decay at the unfolding barrier. (B) p_{fold}^{N}. The error bars represent the standard deviation if the calculations are considered as a Bernoulli experiment. (C) p_{fold} as calculated analytically on the ETN (pfoldt). (D) p_{fold}^{MSM}. The use of the number of native contacts Q for the definition of the folded (Q> 19/26, black circles) and unfolded (Q< 5/26, red circles) state as boundary conditions of the MSM results in incorrect values of p_{fold}^{MSM} and no sharp transition can be observed at the barrier. A similar failure is observed when defining folded and unfolded by rmsd from the native structure (Supp. Mat. Fig. S8).

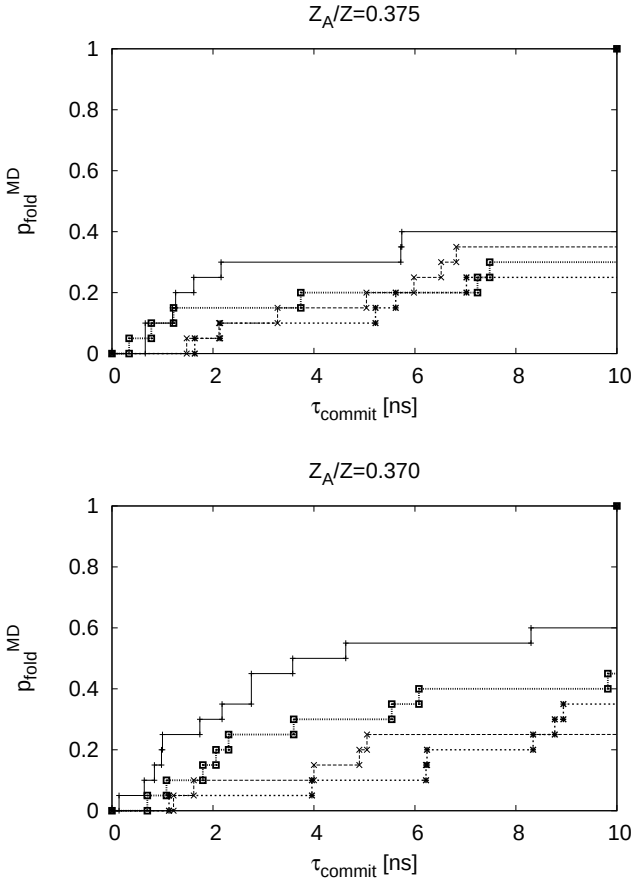


FIG. 4: Dependence of p_{fold}^{MD} on the value of τ_{commit}[15]. Results are shown for 20 short runs from each of four structures of a node with Z_A/Z =0.375 (top) and Z_A/Z =0.370 (bottom). The curves are step functions and reach a plateau at about 5 to 10 ns.

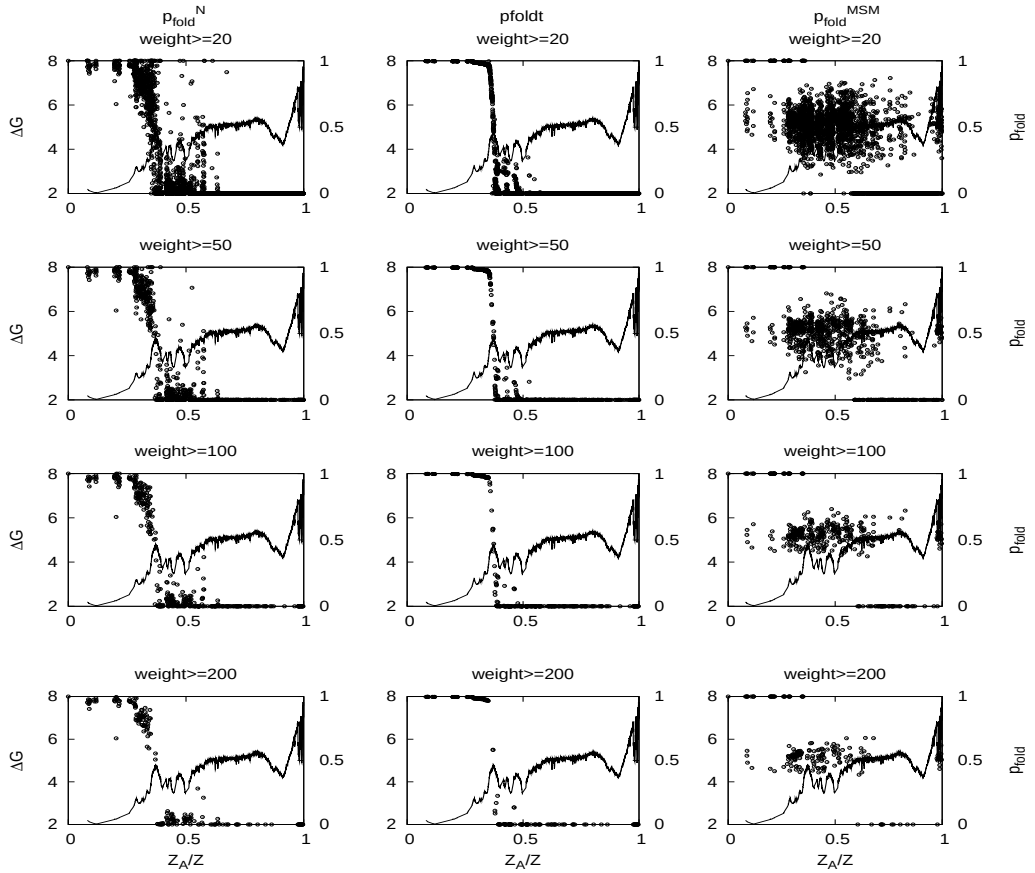


FIG. 5: The cFEP (continuous line, left y-axis) is shown with the values of p_{fold} (empty circles, right y-axis) calculated by p^{N}_{fold} along the trajectories (left), on the ETN (pfoldt, middle), and p^{MSM}_{fold} with Q-based definition of the folded and unfolded ensemble (right). The values are given for all nodes populated by a number of snapshots above a certain cutoff, which increases from top to bottom. Note that pfoldt is the most accurate of the three methods to calculate p_{fold}, and that pfoldt and p^{N}_{fold} improve accuracy for higher weight of the nodes. In contrast, the p^{MSM}_{fold} values are wrong for all node weights because of the boundary conditions defined using Q.

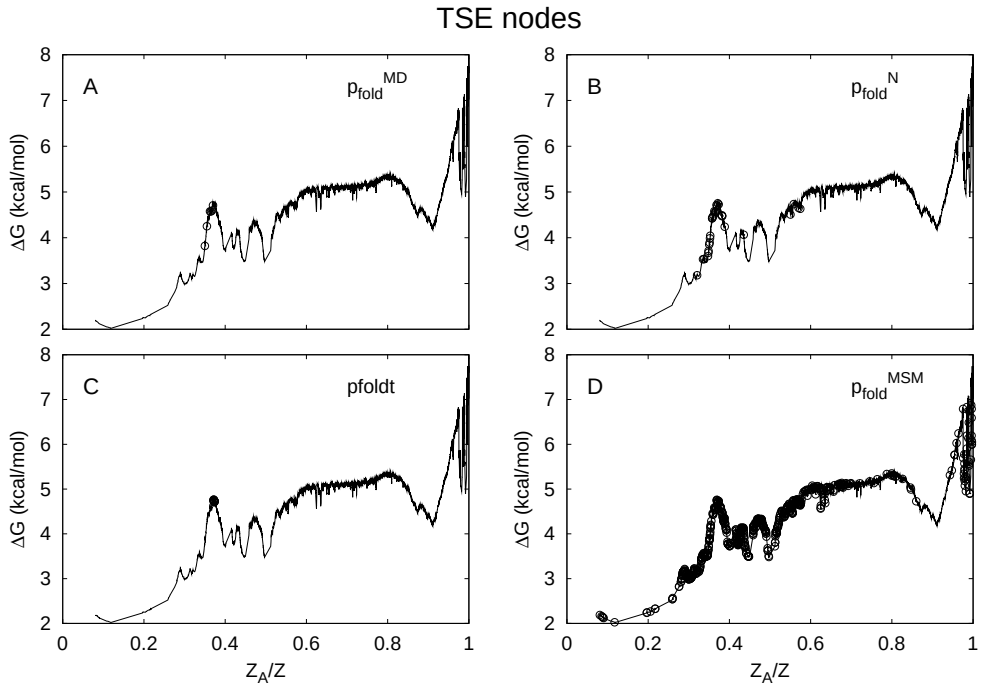


FIG. 6: Putative TSE determined by p_{fold}^{MD} (A), p_{fold}^{N} (B), pfoldt (C), and p_{fold}^{MSM} (D). Nodes with $0.45 < p_{fold} < 0.55$ and 20 or more snapshots are shown (empty circles). (A) Of the 34 nodes used for p_{fold}^{MD} calculations, only four nodes belong to the putative TSE region and are situated close to the top of the cFEP unfolding barrier. (B) Although some nodes with $0.45 < p_{fold}^{N} < 0.55$ are located close to the cFEP unfolding barrier, the TSE isolated by p_{fold}^{N} is affected by statistical error. (C) The TSE isolated by pfoldt is situated exactly on top of the barrier. (D) Most of the putative TSE nodes suggested by the p_{fold}^{MSM} approach do not belong to the TSE.

Identification of the protein folding transition state from molecular dynamics trajectories
SUPPORTING INFORMATION

S. Muff and A. Caflisch

Department of Biochemistry,

University of Zurich,

Winterthurerstrasse 190,

CH-8057 Zurich, Switzerland

tel: +41 44 635 55 21,

fax: +41 44 635 68 62,

e-mail: caflisch@bioc.uzh.ch

Keywords: molecular dynamics; transition state ensemble; p_{fold}; free-energy profile; denatured state ensemble

A. Z_A/Z as progress coordinate

The cFEP projected onto the relative partition function Z_A/Z has the advantage that the first basin on the left (reference basin, usually the folded one) is isolated with its population quantified by the x-axis value at the first barrier on the left. Other progress coordinates can be used, e.g., the mean first passage time (mfpt).The advantage of the cFEP projection onto mfpt is that rates of folding from individual basins are readable from the x-axis[1]. Note that the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate[2].

B. cFEPs with other progress variable than mfpt

The progress *coordinate* (Z_A/Z or mfpt) is used to project the cFEP, while the progress *variable* is required to sort the nodes for the cFEP. For each node in the equilibrium transition network (ETN) two progress variables can be evaluated: mfpt and p_{fold}. Mfpt calculations require the selection of only one node, i.e., the native node[1]. Alternatively, an extra node, which is connected to all nodes in the network by a link weighted proportionally to a Lagrange multiplier λ, is needed in the pfoldf procedure to represent the unfolded state[3]. The introduction of the extra node is a stratagem to circumvent the arbitrary selection of a node as representative of the unfolded state. The results of this work are robust upon the choice of the progress variable (see Figure S3).

C. Supplementary Figures

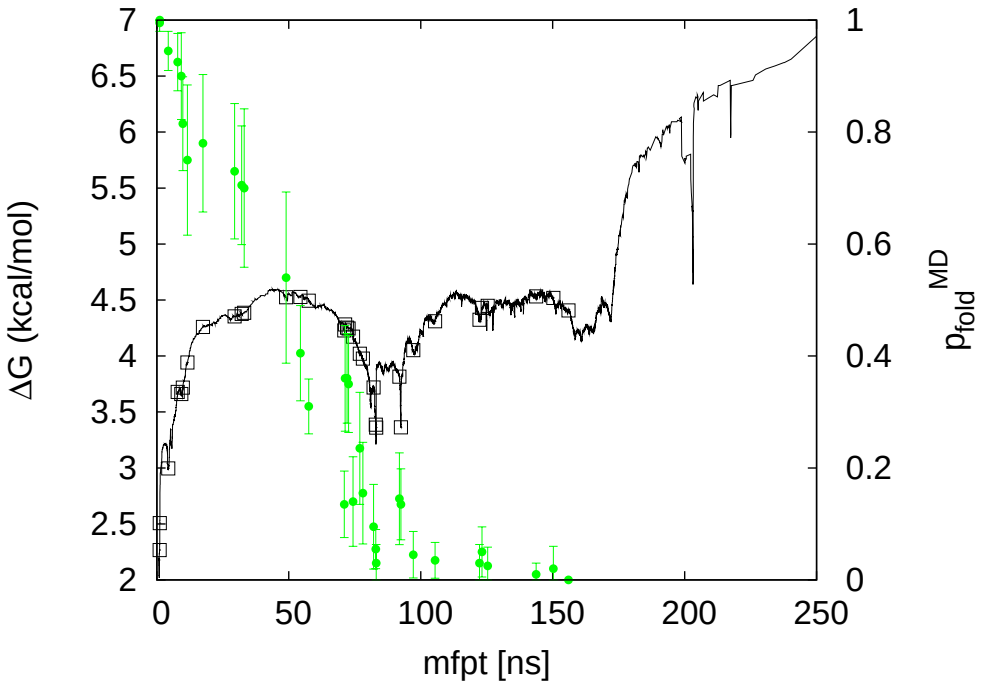


FIG. S1: cFEP with x-axis transformed into mfpt (black line). The p_{fold}^{MD} values (green circles) refer to the right y-axis and are given for the same 34 nodes as in Fig. 3 of the main text (black squares). The decay of p_{fold}^{MD} appears not as sharp as for Z_A/Z because only few nodes populate the region around 50ns, which is the average folding time of TSE structures (because half contribute about 100 ns, and the remaining less than 10 ns)

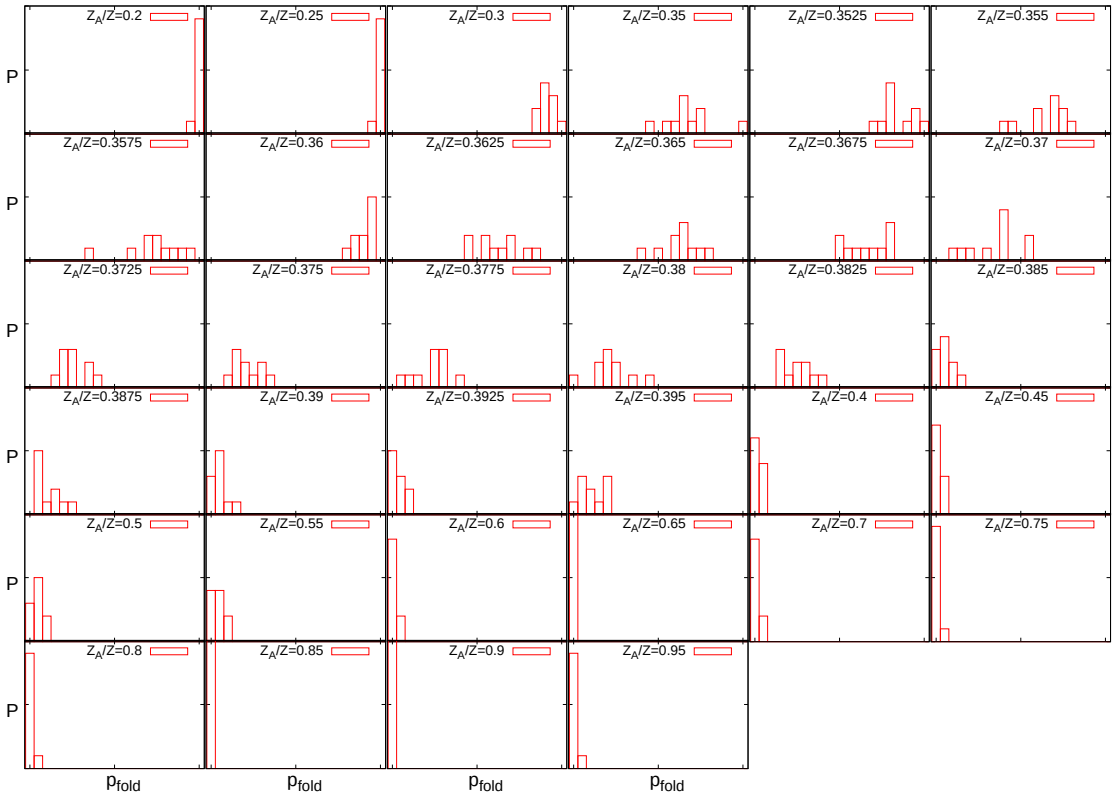


FIG. S2: Normalized histograms of p_{fold}^{MD} for the 34 nodes used for folding simulations. According to these plots, p_{fold}^{MD} values of individual snapshots are peaked around the average value of the respective node, indicating that the coarse-graining procedure applied here groups snapshots in a kinetically homogeneous way.

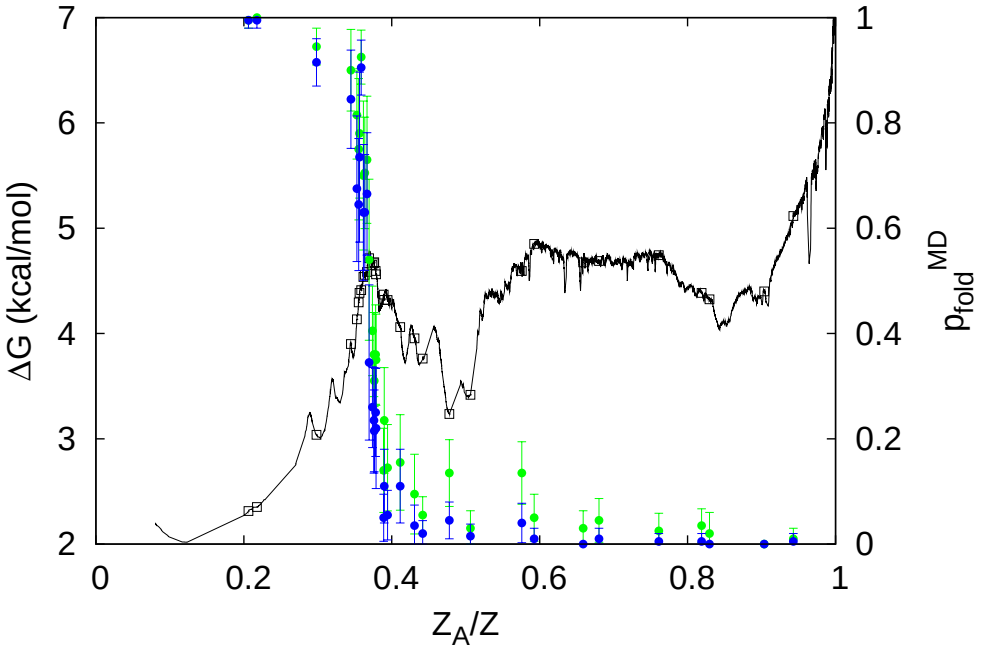


FIG. S3: Pfoldf-cFEP with an extra-node connected by a capacity λ =0.001[3] (black line) and the same selection of nodes as in Fig. 3 of the main text chosen for additional simulations (black squares). p_{fold}^{MD} results for τ_{commit} = 5 ns (blue) and τ_{commit} = 10 ns (green) are essentially identical. The similarity to the corresponding mfpt cFEP (Fig. 3A of the main text) indicates that the procedure is robust upon variation of the progress variable. The similarity of mfpt and p_{fold} profiles is expected, because both encode for kinetic distance to the native state and the equation system for analytical calculation of mfpt and p_{fold} from the ETN differs only in the explicit time dependence of the former[1].

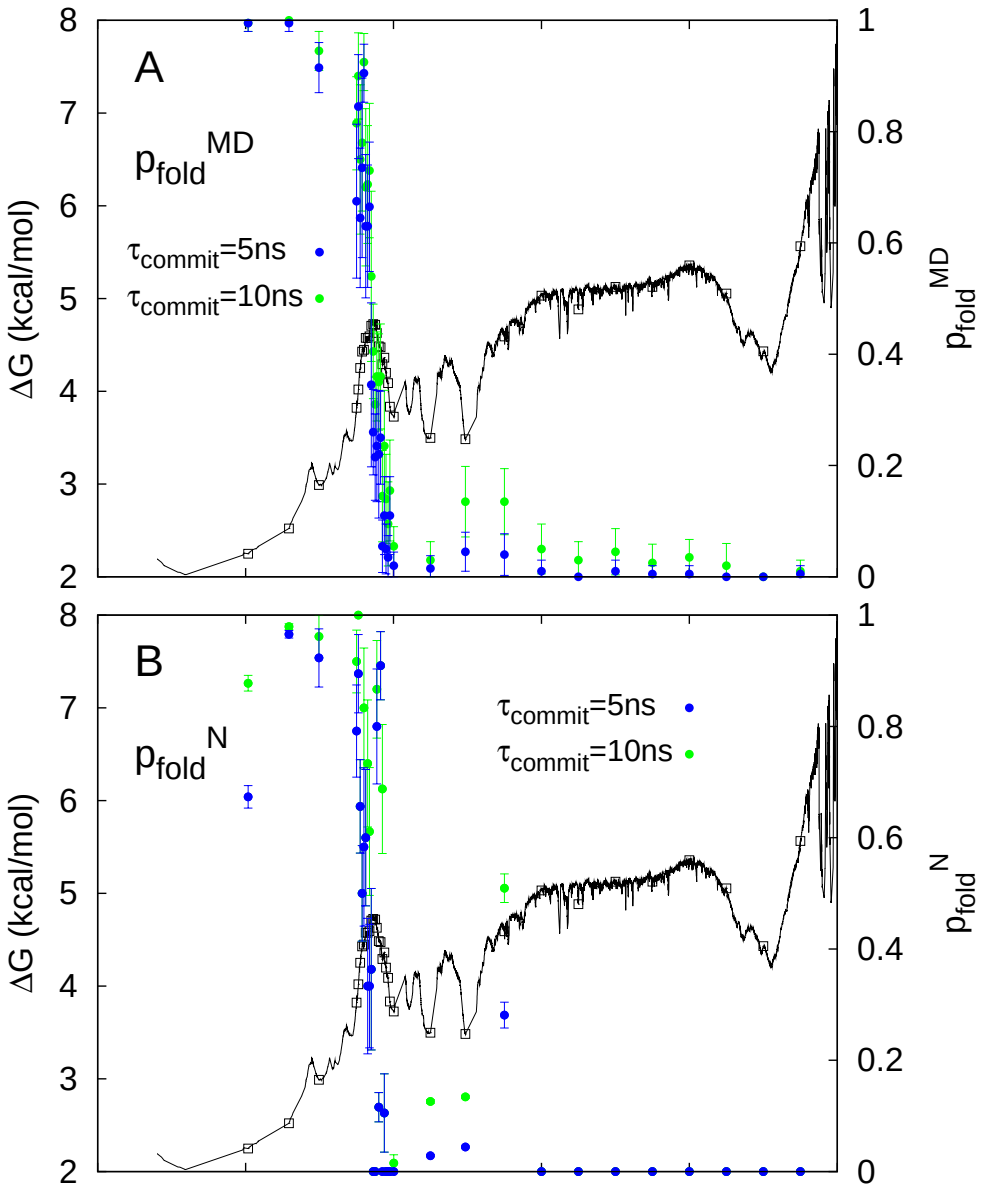


FIG. S4: Dependency of p_{fold}^{MD} (A) and p_{fold}^{N} (B) on τ_{commit}. The same plot as Fig. 3 in the main text, but with $\tau_{commit} = 5$ ns (blue circles) and $\tau_{commit} = 10$ ns (green circles). The results are almost identical.

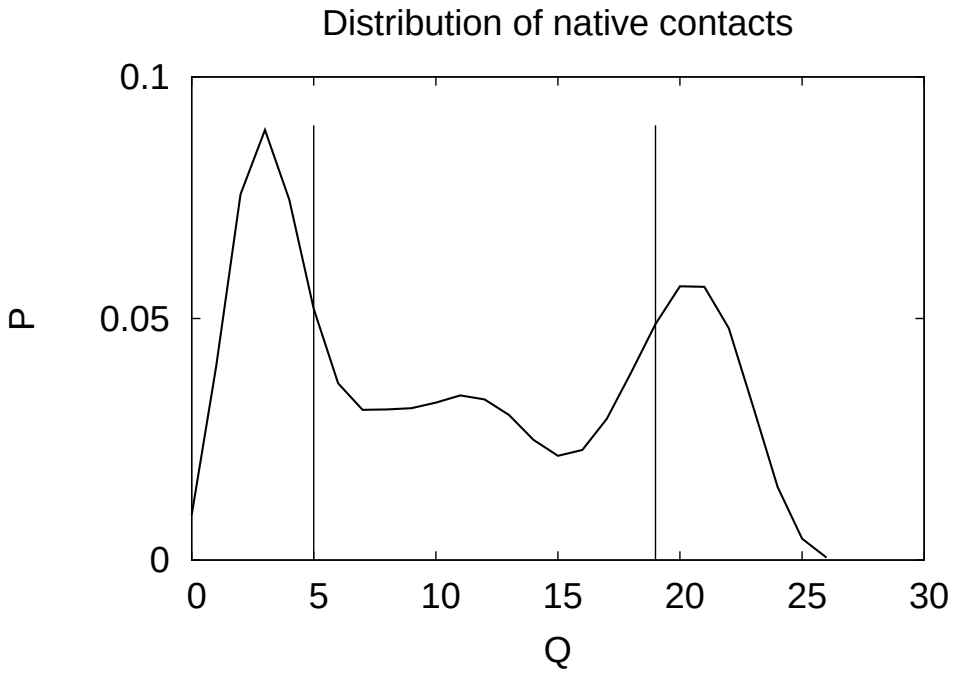


FIG. S5: The 26 native contacts were defined in Ref.[4]. Nodes whose structures have Q > 19 in average were defined as folded, those with Q < 5 as unfolded.

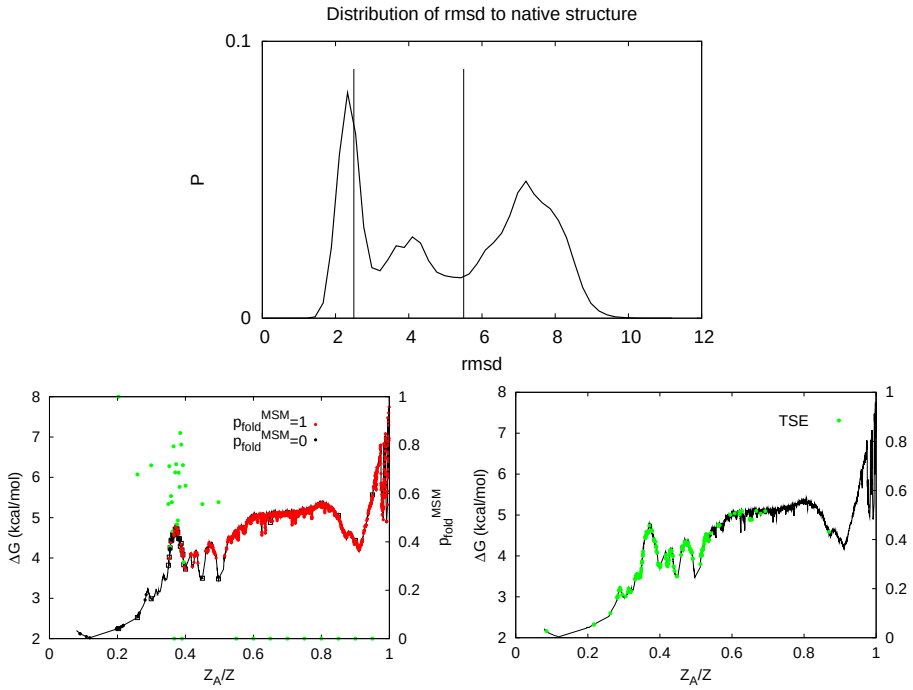


FIG. S6: Results of the Markov state model with rmsd-based definition of boundary states. (Top) Nodes with an average of rmsd < 2.5 Å from the native structure were defined as folded, those with rmsd > 5.5 Å as unfolded. (Bottom) The plots corresponding to Figure 3D (left) and 6D (right) of the main text show that some of the unfolded nodes have $p_{fold}^{\mathrm{MSM}} > 0.5$ and putative TSE structures are suggested far away from the barrier, respectively.

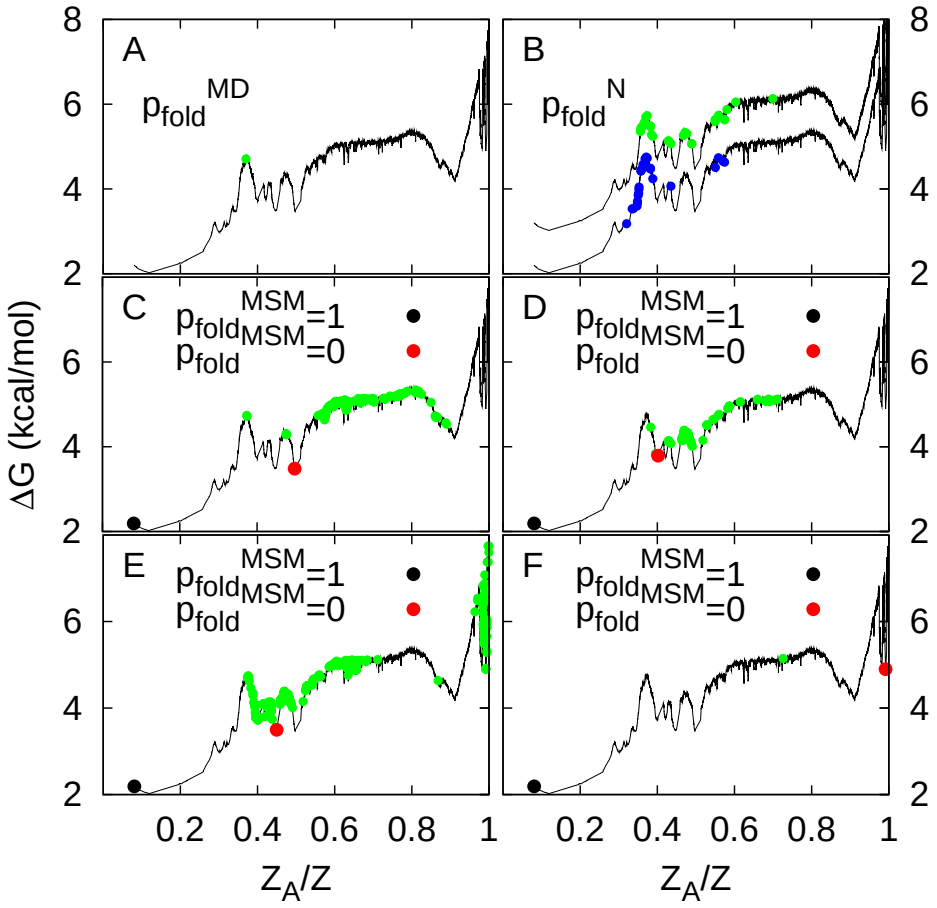


FIG. S7: Correct TSE (A) and putative TSE determined by p_{fold}^{N} (B), and p_{fold}^{MSM} (C-F). Nodes with $0.45 < p_{fold} < 0.55$ and 20 or more snapshots are shown (green circles). (A) Values of p_{fold}^{MD} were calculated using $\tau_{commit} = 10$ ns. (B) Values of p_{fold}^{N} were calculated using $\tau_{commit} = 5$ ns (blue circles) or $\tau_{commit} = 10$ ns (green circles). One of the two profiles is shifted vertically for visual clarity. (C-F) Different representatives of the denatured state (red circles) are used as boundary condition $p_{fold}^{\mathrm{MSM}} = 0$. The profiles are shown to illustrate that most of the putative TSE structures suggested by the p_{fold}^{MSM} approach do not belong to the TSE.

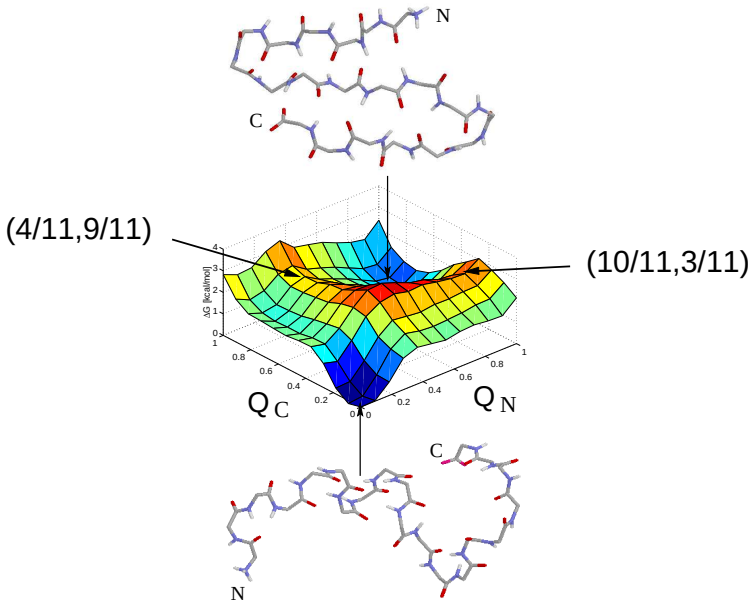


FIG. S8: It is not possible to extract the TSE from conventional histogram-based projections of the free energy onto geometric progress variables. The arrows show the location on the surface of the snapshots used for p_{fold}^{MD} calculations (Figure adapted from [5]).

[1] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).

[2] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. USA. **105**, 13841 (2008).

[3] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).

[4] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. USA. **97**, 10780 (2000).

[5] A. Cavalli, P. Ferrara, and A. Caflisch, Proteins: Structure, Function, and Bioinformatics **47**, 305 (2002).

Chapter 6

ETNA: Equilibrium transition networks and Arrhenius equation for extracting folding kinetics from REMD simulations.

[*Submitted*]

ETNA: Equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations

S. Muff* and A. Caflisch*

Department of Biochemistry, University of Zurich,

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Dated: January 12, 2009)

Abstract

It is difficult to investigate folding kinetics by conventional atomistic simulations of proteins. The replica exchange molecular dynamics (REMD) simulation technique enhances conformational sampling at the expenses of reduced kinetic information, which in REMD is directly available only for very short time scales. Here, we propose a procedure for obtaining kinetic data from REMD by making use of the equilibrium transitions network (ETN) sampled at the temperature of interest. This information is supplemented by mean folding times extracted from ETNs at higher REMD temperatures and scaled according to the Arrhenius equation. The procedure is applied to a three-stranded antiparallel β-sheet peptide which has a very heterogeneous denatured state with a broad entropic basin and several enthalpic traps. Despite the complexity of the system and the REMD exchange time of only 0.1 ns, the procedure is able to estimate folding times (ranging from about 0.1 μs at the melting temperature of 330 K to about 8 μs at 286 K) as well as transition times from individual non-native basins to the native state.

*tel: +41 44 635 55 21 e-mail:caflisch@bioc.uzh.ch,smuff@bioc.uzh.ch

Abbreviations:
ETNA: equilibrium transition network and Arrhenius-equation; REMD: replica exchange molecular dynamics; NC: native component of the ETN; CTMD: constant temperature molecular dynamics; cFEP: cut-based free-energy profile

I. INTRODUCTION

Molecular dynamics (MD) and Metropolis Monte Carlo are simulation techniques widely used for Boltzmann-weighted (i.e., equilibrium) sampling. In principle, the main advantage of MD simulations is the correct description of the dynamics because the time-behavior of the system is not characterized in detail by Monte Carlo sampling [1]. In practice, because of the many degrees of freedom in the (poly)peptide chain and the related complexity of the free-energy landscape it is very challenging to sample the conformational space of peptides and proteins by standard MD techniques, which have an inherently "slow" time step of about 1-2 fs. At low temperatures, MD simulations can get trapped and sample mainly the starting basin. At high temperatures, on the other hand, the accessible phase space increases dramatically and not all possible conformations are visited. A number of simulation techniques have been introduced to enhance the sampling of the conformational space [2–4]. At the same time, the availability of hundreds to thousands of processors has been exploited by intrinsically parallel jobs like distributed computing [5, 6] and loosely coupled MD simulations [7]. Because of the significant time-scale gap between the actual folding process (microseconds to seconds) and simulation length (nanoseconds), it is not possible to extract folding kinetics directly from distributed computing simulations [6, 8]. In this context, Markov chain models have been applied to determine transition probabilities between a small number (usually less than 100) of coarse-grained states from multiple short MD runs [9–11] but the development of an automatic procedure to cluster the MD snapshots into kinetically distinct states is a major obstacle and an active area of research [12–15].

One simulation technique widely used to enhance sampling is replica exchange MD (REMD). In REMD several non-interacting copies of the system are evolved in parallel over a range of temperatures [16]. The values of temperature are exchanged periodically using a Metropolis-like criterion that ensures sampling of the canonical ensemble at each

of these values. REMD is more efficient than constant temperature MD (CTMD) for equilibrium sampling, in particular at low temperatures as shown for peptide folding [17] and aggregation [18]. However, the REMD sampling consists of many discontinuous segments of trajectory, which cannot be used straightforwardly to analyze the kinetics on relevant time scales.

Four approaches to the extraction of kinetics from REMD have been published. Andrec et al. have proposed a network model, in which links represent allowed conformational changes between states according to a geometrical similarity criterion, and each snapshot is a node of the network. Sampling at different temperatures is combined according to the kinetic energy of states [19]. Van der Spoel and Seibert assumed a two-state model and fitted the four parameters of a rate equation by employing the fraction of native folded species along a heterogeneous set of 16 REMD and 4 CTMD trajectories of a β-hairpin decapeptide [20]. Yang et. al approximate the folding process by Langevin dynamics along a one-dimensional reaction coordinate R with effective random forces and diffusion coefficient (as a function of R) extracted from REMD [21]. Their approach requires the *a priori* definition of a one-dimensional reaction coordinate for folding which almost always masks the hidden complexity of the folding process [22–24]. This complexity is also masked in the two-state assumption of van der Spoel and Seibert. Recently, Buchete and Hummer have proposed a procedure to extract rates from the number of transitions, on the time scale of replica exchanges, by calculating the rate coefficients of a master equation using the maximum likelihood technique [14, 25]. They applied this procedure to the blocked alanine pentapeptide in explicit water (which was coarse-grained into 32 states according to a 5-bit string of residue helicity) but concluded their letter by explicitly mentioning that the application to protein folding might "pose a major challenge" because of the large number of states [14].

Here, we present a procedure for extracting kinetics from REMD which can be applied to systems more complex than those mentioned above, e.g., peptides and proteins simulated at atomistic resolution. First, the equilibrium transitions network (ETN) is constructed for each value of the temperatures used in REMD. More precisely, the ETN is the capacitated graph whose nodes and links represent coarse-grained microstates and transitions, respectively, sampled in the short segments at constant temperature. The ETN often consists of several disconnected components because of the short trajectory

segments between replica exchanges and due to free-energy barriers separating states. Within each component an equilibrium phase-space distribution at the respective temperature is sampled because of the canonical-ensemble sampling within the REMD segments. An important aspect of the procedure for extracting kinetics from REMD is that the ETN can be treated as a Markov state model, implying that Monte Carlo simulations on the network reproduce the correct dynamics within each component. To estimate folding rates at each temperature, mean folding times (mfts) are computed as in [26, 27] on the ETN component that is connected to the native state. Finally, the Arrhenius equation and the sampling at high temperatures are used to extract kinetics for the low temperature nodes that are disconnected from the NC of the ETN (Figure 1). The procedure is termed ETNA because of the combination of the ETN and the Arrhenius equation. Moreover, thanks to the thermodynamically correct sampling from the REMD trajectories and the integration of short REMD segments into ETN components, it is possible to extract correct populations of enthalpic free-energy basins from the analysis with cut-based free-energy profiles (cFEPs), which is a method for grouping conformations according to (local) transitions at equilibrium [28].

ETNA is applied to the miniprotein called Beta3s [29] whose native structure corresponds to a three-stranded antiparallel β-sheet consisting of two β-hairpins [30]. Beta3s has been shown to fold to the native structure determined by NMR [30] in molecular dynamics simulations with the CHARMM polar hydrogen molecular mechanics potential energy function supplemented by a simple implicit solvent model [29]. Since folding simulations of Beta3s are very fast close to its melting temperature of 330 K (folding time of about 0.1 μs, which requires roughly 18 hours on a single core of a XEON 2.33 GHz), many studies have been made to elucidate its folding mechanism [15, 17, 23, 27, 29, 31, 32]. ETNA is able to extract from REMD overall folding times of Beta3s, as well as folding times from individual basins in the unfolded state, that are in good agreement with the corresponding values obtained by multiple CTMD folding runs started from the denatured state ensemble at 286 K. Therefore, kinetics on time scales five orders of magnitude longer than the REMD segments are accessible, as the REMD exchange time was only 0.1 ns and the folding time of Beta3s is about 8 μs at 286 K.

II. THEORY

A. Equilibrium transition network (ETN) from REMD segments at constant temperature

The trajectory segments collected in REMD simulations at a given temperature are much shorter (picoseconds) than the time scales of large conformational transitions or folding (microseconds to seconds). The length of the segments depends on the frequency of the swapping attempts and their acceptance ratio, which is usually 25-30%. These segments of trajectory are between three and six orders of magnitude shorter, depending on the temperature, than the folding time of a structured peptide or a small protein. The essential idea of ETN is to extract kinetics from the integration of all REMD segments at the same temperature. For complex systems, the ETN at each temperature consists of several disconnected parts, one of which contains the native state and is termed native component (NC) hereafter. The NC is usually the largest component, but its size can be reduced at very low temperature due to large free-energy barriers between states, as well as at high temperature (above 307 K for Beta3s, see Results) because sampling is not sufficient to fully connect the large accessible space, especially in the presence of entropy-dominated regions.

There are two important conditions on the ETNs. First, the individual ETN components must fulfill the property of Markov state models. Markovianity depends on the way how snapshots are grouped into nodes and on the lagtime of the transitions. The second condition is that all components represent *locally* the correct connectivity and population of states, i.e., that the ETN assembled from REMD sampling is indistinguishable from the corresponding portion of the ETN from converged CTMD simulations. This requires that the REMD exchange time is long enough for establishing local connectivity.

Note that for an ETN generated by the combination of thousands of short trajectories, like in REMD, it is important to symmetrize the transition matrix (i.e., impose detailed balance) by replacing the absolute number of transitions n_{ji} from node i to node j by $c_{ji} = \frac{n_{ji}+n_{ij}}{2}$. Such an enforced detailed balance is allowed only if the REMD simulations are long enough to reach equilibrium at all temperatures. While this step is helpful (but not essential) for long equilibrium trajectories [28], it is necessary for ETNs extracted

from REMD to avoid dead-ends. A dead-end may arise when the trajectory is interrupted because of a temperature swap, leaving the last visited node without a next neighbor. Such nodes are problematic if transition times are calculated by solving the respective master equation on the ETN (see next subsection).

B. Mean folding time calculation on the NC at constant temperature

The mean folding time (mft) is the mean first passage time to the native node. Given the transition probability p_{ij} between nodes j and i (with $p_{ij} = c_{ij}/\sum_k c_{kj}$), the mft for a node i in the NC at a given temperature is the solution of the equation system $\mathrm{mft}_i = \Delta t + \sum p_{ij} \cdot \mathrm{mft}_j$, which can be determined by iterative multiplication [26, 27]. Δt is the lagtime time of the Markov state model. Solving the equation system allows the calculation of the mft from nodes in the NC that actually never fold within any of the short REMD segments. On the other hand, it is not possible to calculate the mft of nodes not belonging to the NC. For snapshots in non-NC nodes, the Arrhenius-scaling approach (ETNA) is introduced as follows in the next subsection.

C. Scaling folding times using the Arrhenius equation

An essential aspect of the ETNA procedure is the use of the Arrhenius equation and high-temperature sampling to extract kinetics at low temperature for microstates that do not belong to the NC. Assuming both the pre-exponential factor A and the activation energy to exit from a minimum of interest E_a to be temperature independent, the ratio of folding rates from the respective basin k_i at different temperatures T_1 and T_2 is

$$\frac{k_2}{k_1} = \frac{Ae^{-\frac{E_a}{RT_2}}}{Ae^{-\frac{E_a}{RT_1}}} = e^{\frac{E_a}{R}(1/T_1 - 1/T_2)}$$

$$\Rightarrow \quad \tau_1 = \tau_2 \cdot e^{\frac{E_a}{R}(1/T_1 - 1/T_2)} \; . \tag{1}$$

In a first approximation, E_a/R can be taken as a universal constant of the system and extracted by a linear fit of the $1/T$ vs. $ln(k)$ plot of unfolding rates at several temperatures. Note that this assumption of universality for E_a/R is invalid if folding barriers

from different non-native regions are very heterogeneous or very different from the unfolding barrier, but a more general theory with multiple scaling factors can be derived in a straightforward way.

The Arrhenius equation is an approximation that ignores entropic contributions. Using the Eyring equation from transition state theory, the ratio of reaction rates can be written as

$$\begin{aligned}\frac{k_2}{k_1} &= \frac{T_2}{T_1} \cdot e^{(\frac{-\Delta G}{RT_2} - \frac{-\Delta G}{RT_1})} \\ &= \frac{T_2}{T_1} \cdot e^{(\frac{-\Delta H + T_2 \Delta S}{RT_2} + \frac{\Delta H - T_1 \Delta S}{RT_1})} \\ &= \frac{T_2}{T_1} \cdot e^{(\frac{-\Delta H}{R}(\frac{1}{T_2} - \frac{1}{T_1}))} \ .\end{aligned}$$

Thus, under the simplifying assumption that ΔH and ΔS are temperature independent, the entropic contribution $T\Delta S$ cancels in the ratio of rates even when the Eyring approach is used. Moreover, the "pre-factor" T_2/T_1 is close to 1.0 for similar temperatures, as they are usually employed in REMD simulations. Therefore, the use of the (simpler) Arrhenius equation is justified.

As mentioned above, the scaling of folding times according to the Arrhenius equation comes into play because at low temperatures the ETN from REMD is usually split into disconnected pieces due to high free-energy barriers that separate basins (Figure 1). Therefore, folding times from outside the NC at a low temperature of interest (T_1) cannot be calculated directly on the ETN. When the temperature of a replica is swapped to the next higher temperature (T_2) in the REMD simulation, the trajectory moves to the ETN at T_2. If nodes of the ETN at T_2 are visited, the mft of the closest (in time) node is scaled to T_1 according to the Arrhenius equation (1) and the snapshots in the previous T_1 segment are assigned an mft. If the procedure is not successful for T_2, the next temperature T_3 is considered and so on. If between two T_1 segments the system does not visit the NC at any other temperature, it is not possible to assign mfts to the previous T_1 segment. Those snapshots remain unassigned and are therefore ignored. Note, however, that the scaling of folding kinetics with ETNA is valid only in temperature ranges where

folding times follow the Arrhenius law. If a temperature T_A is known, where the system starts to show an anti-Arrhenius behavior, the data with $T > T_A$ must be discarded from the analysis. Hence, the snapshots in a T_1 segment are ignored if no NC at a temperature between T_1 and T_A is visited before the system continues to $T > T_A$.

Figure 1 illustrates the ETNA algorithm for the case where nodes at T_3 are used to scale a fragment of the trajectory at T_1. The mft of the first of these T_3 *nodes* is used to calculate the theoretical mft of the last T_1 *snapshot* (mft_1), taking into account also the effective time τ_2 spent in the segment at T_2, $\text{mft}_1 = \text{mft}_3 \cdot e^{\frac{E_a}{R}(1/T_1 - 1/T_3)} + \tau_2 \cdot e^{\frac{E_a}{R}(1/T_1 - 1/T_2)}$, where mft_3 was previously calculated by solving the system of equations at T_3. Since a snapshot cannot have a mean folding time, but only one value originating from one folding event along the trajectory, the folding time τ assigned to the last T_1 snapshot is chosen randomly according to the exponential distribution around mft_1 as $P(\tau) = k \cdot e^{-k\tau}$ with $k = \frac{1}{\text{mft}_1}$. This last step is essential in order to obtain, in addition to the average value, a cumulative folding time distribution, which is used later for analysis. All remaining T_1 snapshots in the considered segment are assigned a folding time exponentially distributed around $\text{mft}_1 + i \cdot \Delta t$, with i being the number of timesteps backward from the last snapshot in the segment, and Δt the lagtime of the model. Note that with this procedure the mft scaling from higher temperature to the reference temperature is done separately for every snapshot in nodes not connected to the NC of T_1. Therefore the ETNA approach is different from pure Arrhenius-based methods [20], because each snapshot is assigned an individual folding time value, which depends on the route the system takes for folding.

D. Cut-based free-energy profiles (cFEPs)

The cFEP approach was first introduced in [28] and further developed in [27]. For a node i in the ETN the partition function is $Z_i = \sum_j c_{ij}$ where, as mentioned above, c_{ij} is the symmetrized number of transitions between nodes j and i. If the nodes of the network are partitioned into two groups A and B, then $Z_A = \sum_{i \in A} Z_i$, $Z_B = \sum_{i \in B} Z_i$, $Z_{AB} = \sum_{i \in A, j \in B} c_{ij}$ and the free energy of the barrier between the two groups is $-kTlog(Z_{AB}/Z)$ with Z being the partition function of the full network (Figure 2). The cFEP has the advantage with respect to the projections onto geometric coordinates that barriers are preserved [28]. In particular, the relative partition function Z_A/Z includes all pathways

to and from the state of interest (e.g., the folded state). The cFEP method groups conformations according to equilibrium kinetics. Their application to components of the ETN is possible, because transitions from constant temperature segments establish locally the correct connectivity. Therefore, the profiles of ETN components are expected to be identical to those that would be extracted from equilibrium sampling. The cFEP analysis was performed with the program WORDOM [33], which is particularly efficient in handling large sets of trajectories. Here, only cFEPs with mft as progress variable are used. Values of mft for individual nodes are calculated as explained above.

E. Isolation of free-energy basins

Since the ETN constructed from REMD segments at constant temperature yields multiple disconnected components, it is not possible to obtain the complete cFEPs, i.e., the profile up to $Z_A/Z = 1$. However, the majority of nodes within a given free-energy basin belong to the same component of the ETN, at least if relaxation in the basin is as fast as the minimal length of the segments, which ensures that different REMD segments are connected through their visits to some of the highly populated nodes. Therefore, the procedure to extract basins from the cFEP remains the same as for the NC, where unfolding cFEPs from a node in the basin of interest (usually its most visited node) are plotted. The nodes lying on the left of the cut at the first barrier make up the basin.

III. APPLICATION OF ETNA TO BETA3S

A. Molecular dynamics simulations

All simulations and part of the analysis of the trajectories were performed with the program CHARMM [34]. The designed 20-residue peptide Beta3s [30] was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field [35] with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent on the solute [36]. It was shown previously using exactly the same SAS-based implicit solvent model that at 330 K Beta3s folds reversibly to its NMR conformation irrespective of the starting structure,

and importantly, 23 of the 26 NOE restraints are satisfied [29]. Despite the absence of collisions with water molecules, in the simulations with implicit solvent relative rates of folding of structured peptides are comparable with the values observed experimentally [37–39]. Importantly, the small variations in total SAS and radius of gyration during folding of Beta3s at 330 K [31] suggest that the lack of solute/solvent friction does not have a significant effect on pathways and kinetics.

B. REMD setup

In the present simulations, eight replicas were run with temperatures (in K) of 286, 307, 330, 355, 382, 411, 442, and 476 for a simulation time of 11 μs each. Swapping attempts between replicas were performed every τ_{swap}=0.1 ns with an acceptance ration of about 25% and thus most REMD segments are 0.1-1 ns long. The Berendsen thermostat was used with a much shorter coupling constant of 5 ps to allow the temperature of the system to relax between two swapping attempts. Frames were saved with a frequency of 20 ps and therefore a REMD segment contains at least 5 consecutive snapshots before a new temperature is accepted. The low swapping- vs. saving-frequency was chosen in order to let the system sample local transitions, which are the essential ingredient in the method presented here.

C. CTMD folding runs at 286 K and 307 K

It is computationally prohibitive to obtain reversible folding-unfolding of Beta3s by CTMD at low temperature. Therefore, 750 CTMD folding runs at 286 K and 250 at 307 K were performed for comparison with REMD. Starting conformations were chosen uniformly distributed over the denatured state ensemble in the REMD segments at 286 K and 307 K (see subsection II E for definition of the native basin). Folding is defined by all-atom rmsd $\leq$ 2.5 Å from the snapshot in the center of the folded node in the REMD sampling at the respective temperature, as identified by the leader algorithm [40]. Therefore, a folding event is defined through the same structural constraints in both the CTMD folding runs and REMD. The CTMD simulations were stopped upon folding or after 10 μs, even if the folded state was not reached because of the large computational

cost (about 300 days on a 200-CPU cluster). Note that 164 out of the 750 and 28 out of the 250 CTMD runs at 286 K and 307 K, respectively, did not fold within 10 μs. Nevertheless, the 10 μs could be included in the cumulative folding time distribution $f(t) = \int_t^\infty p(\tau)d\tau$, because $f(t)$ is the probability that a folding event requires *at least* time t.

D. The Markov state model of Beta3s

It is necessary to coarse-grain the snapshots because each conformation is visited only once; in other words, any trajectory, per se, is nothing but a long string of coordinate sets. There are several meaningful ways for clustering individual coordinate sets in the trajectory to obtain coarse-grained microstates (nodes is used a synonymous in this paper), and different ones are likely to be most useful for different types of analysis. For a structured peptide like Beta3s or a β-hairpin, rmsd and secondary structural coarse-graining are obvious possibilities [22, 23, 41]. The coarse-graining used in this work is the leader algorithm based on the all-atom rmsd [40] with a cutoff of 2.5 Å. Note that nodes in the ETN with only one or two neighbors (i.e., one incoming and/or one outgoing neighbor) were grouped to their outgoing neighbor. This regrouping is justified because the future of such nodes within a trajectory is completely determined, i.e., no information is erased through their regrouping. Upon rmsd coarse-graining and regrouping the following numbers of nodes were found: 4'183 (at 286 K), 11'611 (307 K), 26'719 (330 K), and 38'445 (355 K). These nodes are the states of the Markov state model (i.e., the ETN), and the lagtime was set to Δt=20 ps. Figure 3 contains a comparison of folding dynamics from the CTMD simulations and from the corresponding Markov state model at 330 K. There is a very good agreement for the overall dynamics, as well as for folding distributions from various metastable states, indicating that the Markov assumption holds.

E. The Arrhenius fit

As explained in the Methods section, the parameter E_a/R for the fit of the Arrhenius plot can be calculated from unfolding rates at different temperatures (Figure 4). The simplest way to obtain rates is by estimating them from the ETN of a CTMD simulation.

Unfolding rates were extracted as the average time the system spends in the folded state before exiting from it, where the folded state is defined from the cFEP of the considered temperature by cutting at the first significant barrier in the profile (Figure 5) [27, 28]. Unfolding rates extracted with this procedure are shown as squares and fitted by the solid line in Figure 4a, from which the slope $E_a/R = 6730$ K was extracted.

It is also possible to approximate unfolding rates directly from the REMD data. This approach can be especially useful if equilibrium or unfolding simulations are too expensive, e.g., for large systems or if the unfolding barrier is very high. The procedure to estimate the rates is the same as for CTMD, with the only difference that the ETN is constructed from REMD (and not from CTMD) data and only the NC can be used. Figure 4a contains the rates estimated on the CTMD and REMD ETNs, where the slope $E_a/R = 9640$ K is obtained by fitting the latter. Figure S2 in Supp. Mat. shows that the main results obtained by ETNA are robust with respect to the type of simulations used to extract the value of E_a/R, implying that the REMD data is sufficient and no costly CTMD simulations to estimate unfolding rates are needed.

Interestingly, rates to exit other enthalpic basins can be fitted with similar E_a/R (Figure 4b and c), which means that the differences in activation energy to leave enthalpic basins of Beta3s are relatively small. Therefore, the approximation of using the activation energy for unfolding as a representative barrier to leave any enthalpic basin of the system is valid in the application of the ETNA procedure to Beta3s.

F. cFEPs from REMD data at individual temperatures

The profiles from the ETN at each of the four lowest REMD temperatures are shown in Figure 5. These cFEPs represent only the NCs of each ETN, and include the indicated portions of the sampled conformational space. A comparison of the same profiles with the results from CTMD simulations at 330 K, 355 K and 382 K shows a remarkable similarity up to the first significant barrier (Supp. Mat. Figure S6), which indicates that the ETN from REMD sampling contains indeed the correct connectivity information.

At 286 K and 307 K the main contribution originates from the native basin (with a weight of 60.6% and 54.3%, respectively). At higher temperatures the native state shrinks. Only 37.8%, 16.8%, and 1.3% remain native at 330 K, 355 K, and 382 K,

respectively. Thus, even though the absolute size of the NC decreases for temperatures above 307 K, more non-native basins belong to the NC with increasing temperature.

G. Removal of entropic effects

Beta3s is known to spend about one third of its time in an entropic region at 330 K, i.e., in a non-native region with heterogeneous structures stabilized mainly by entropy [27]. Even in a 20-μs equilibrium simulation at 330 K this entropic region suffers from incomplete sampling and the majority of the nodes is visited only once or a few times [15, 23]. The entropic fraction increases dramatically at higher temperatures. The insufficient sampling of these regions introduces large errors to the ETN. Therefore, most parts of the entropic regions were ignored for calculations on the ETNs and only the well sampled portion, corresponding to the enthalpic basins, was used to be consistent with the Arrhenius equation, which is valid for enthalpic barriers. Figure 5 shows how this selection was carried out for different temperatures with the help of the information from the cFEPs. At 286 K and 307 K no entropic contribution is present and all nodes were considered. At 330 K and 355 K, the cFEPs (solid lines in Figure 5) show pronounced minima which represent the enthalpic basins. After the last enthalpic basin along the Z_A/Z coordinate, the cFEPs are clearly entropy-dominated, as can be seen from their rough shape which indicates insufficient sampling. All nodes above the threshold indicated in the profiles were discarded. The removal of these regions at high temperature does not bias the scaling of the kinetics by ETNA from high to low temperature, since a very small part of the free-energy surface is entropic at low temperature. In addition, temperatures higher than 355 K were not used in the application of the ETNA procedure to Beta3s, because there the folding rate shows a clear anti-Arrhenius transition according to Figure 6, i.e., T_A (see subsection II A) was chosen as 355 K.

H. Free-energy basins

All significantly populated free-energy basins with enthalpic stabilization could be determined from different ETN components. cFEPs from various basins that belong to different components at 286 K and 330 K are given in Supp. Mat. Figures S4 and S5,

respectively. States at different temperature are considered to correspond to each other if the most populated DSSP secondary structure string [43] (first column in Table I) is the same. This comparison ensures that the bottom of the corresponding basins contain similar conformations, but it clearly does not imply that the basins are completely identical and such an assumption is not used anywhere in this work. Populations extracted from the cFEPs are presented in Table I. At 330 K the thermodynamics can be compared to those from a 20-μs equilibrium CTMD simulation. The results are in high agreement, except for the Ch-curl$_1$ enthalpic trap, which was visited only once in the CTMD trajectory and therefore has a large error. The high agreement between CTMD and REMD thermodynamics, both extracted by the cFEP approach, is not trivial because the cFEP method is based on the information of the *equilibrium* transitions between states, whereas REMD samples the correct ensemble of conformations, but only *local* transitions. Therefore, the use of cFEPs is only possible if transitions at constant temperature are sampled, as it is the case here because the REMD swapping frequency was chosen lower than the saving frequency of conformations.

I. Folding time estimates from REMD

The cumulative folding time distribution from nodes outside the native basin at 286 K is shown in the top left panel of Figure 7. The red control distribution from the 750 CTMD folding runs can be fitted between 1 and 10 μs with $e^{-t/7.76\mu s}$. Within the same interval, folding kinetics extracted from REMD with the ETNA procedure scale almost identically as $e^{-t/7.78\mu s}$. As a comparison, if only the non-native part of the NC is used, i.e., if the Arrhenius-scaling is not applied, the ETN of REMD would suggest a folding time of roughly 0.7 μs and therefore underestimates the real folding kinetics by one order of magnitude. Note that the ETNA procedure is able to scale only folding times from a fraction of all snapshots outside of nodes from the NC, because if no NC-node from the network of a temperature between T_1 and T_A is visited before the replica continues to $T > T_A$ (see Methods), all snapshots of the previous T_1 segment are ignored. At 286 K only 20.5% of the snapshots from nodes outside the NC could be assigned a folding time with ETNA. This result implies that the scaling of even a small fraction of folding kinetics is sufficient to yield correct overall rates.

At 330 K (Figure 7, bottom left), the CTMD kinetics were fitted for values up to 600 ns with $e^{-t/158ns}$, while the ETNA-scaled times are only moderately faster ($e^{-t/143ns}$). The folding times for trajectories starting from the ETN (i.e., only considering the NC) are distributed as $e^{-t/134ns}$, thus unlike at 286 K, the application of the Arrhenius-based scaling of rates from different temperatures has almost no effect at 330 K. Similar cumulative folding time distributions are obtained by ETN and ETNA because at higher T the non-native regions of the NC are significantly populated, which reduces the effect of the Arrhenius scaling. According to Figure 7, the NC-ETN at 355 K and even the one at 307 K are sufficient to reveal approximately correct folding rates. Note that, since folding times from high temperatures are used to scale rates at low temperatures with ETNA, the availability of correct folding times from at least one higher temperature is necessary to obtain correct folding kinetics at low temperature.

In addition to overall folding time distributions, kinetics from individual basins were estimated at 286 K and 330 K (Table I). For basins belonging to the NC at the respective temperature, it is not necessary to use Arrhenius scaling to estimate folding times because values of mft can be calculated directly on the NC. In contrast, the mft of basins not belonging to the NC have to be evaluated with the Arrhenius approach. Due to the assignment of folding times to individual snapshots with ETNA, often even only a portion of all snapshots belonging to nodes of a basin can be scaled. This problem is severe in the case of Ch-curl$_1$, for which less than 5% could be assigned a folding value. In such a case the folding time estimate is very inaccurate and it can be helpful to plot the cumulative folding time distribution. The latter does not contain all details of the folding kinetics, but is in return less sensitive to noise and statistical errors than plain distributions [42]. Therefore, the exponential fit to the former was used to estimate the folding kinetics from all basins (Supp. Mat. Fig. S5).

Similarly, the statistics harvested for individual basins from the 286 K CTMD folding runs are relatively low, because starting points of the 750 runs were distributed over all basins. Nevertheless, deviations between folding times from individual basins obtained with REMD or CTMD might originate from low statistics, yet the values lie within the same order of magnitude. The exception is Cs-or at 286 K, which exemplifies the main caveat of the ETNA approach. The Arrhenius equation approximates only the enthalpic contribution of barriers. Therefore, the folding time scaling of entropically

stabilized regions of the free-energy landscape is not valid. Due to the considerable entropic stabilization of the Cs-or basin, which was reported earlier [27] and emerges also from a comparison of its statistical weight at 286 K (0.7%) and 330 K (5.3%), the scaling of the folding time at 286 K overestimates by one order of magnitude the mft from the CTMD simulations (Table I).

IV. CONCLUSIONS

ETNA (equilibrium transitions network and Arrhenius scaling) is a procedure to extract kinetics from REMD simulations. At each of the REMD temperatures, the procedure makes use of the network whose nodes and links are the clustered snapshots and the transitions observed in the short REMD segments, respectively. These networks consist usually of a component that includes the native state and several disconnected components. An essential element of ETNA is the use of the Arrhenius equation for scaling mean folding times of nodes at temperature values higher than the temperature of interest. In this way, folding times at the latter temperature can be estimated for the nodes that are not connected to the native component. The use of the Arrhenius equation is the main difference between the ETNA procedure and a previously published approach based on the distribution of the kinetic energy [19].

There are three conditions to apply the ETNA procedure. First, each component must fulfill the properties of a Markov state model. Second, the REMD segments should be long enough (i.e., the temperature-swapping frequency low enough) to allow for local transitions to take place at constant temperature in REMD, so that the ETN components at each REMD temperature are *locally* indistinguishable from the ETN obtained by a long CTMD simulation.. Third, it is assumed that the scaling in terms of the Arrhenius equation is appropriate, i.e., the free-energy basins are mainly enthalpic, so that the mean folding rate of a node is essentially identical to the corresponding rate constant for the entire basin. However, folding rates from different basins do not necessarily have to be identical and an adaptive scaling approach might be derived in the future.

ETNA was applied to extract folding kinetics at low temperature from a REMD simulation of Beta3s, a three-stranded antiparallel β-sheet peptide of 20 residues. Beta3s is a challenging test system because of its complex denatured state, which consists of sev-

eral enthalpic traps, a basin with fluctuating helical conformations, and a heterogeneous entropic region at temperature values close to the melting temperature. Notably, overall folding rates of Beta3s and folding times from mainly enthalpic non-native basins are estimated correctly by ETNA. Moreover, the folding time of about 8 μs at 286 K is in agreement with NMR data (4-14 μs at 283 K) [30].

We plan to apply ETNA to extract folding kinetics of small proteins simulated by REMD with an efficient and accurate implicit solvent model [44]. Moreover, ETNA can be employed to investigate the kinetics of other biologically relevant processes like large conformational transitions involved in enzyme or receptor functions.

Acknowledgments

We thank Sergei V. Krivov, Philipp Schütz, and François Marchand for useful comments to the manuscript. The simulations were performed on the Etna and Matterhorn clusters of the University of Zurich and we thank Christian Bolliger for hardware support. This work was supported by a Swiss National Science Foundation grant to A.C. Procedures for calculating the cut-based FEPs are available in WORDOM `http://www.biochem-caflisch.uzh.ch/wordom`.

Supporting Information Available

Cumulative folding time distributions with E_a/R=9640 K and additional cFEPs are shown in the Supplementary Material. This information is available free of charge via the Internet at http://pubs.acs.org.

[1] M. Karplus and J. A. McCammon, *Nature Struct. Biol.*, **2002**, *9*, 646–652.

[2] B. J. Berne and J. E. Straub, *Curr. Opin. Struct. Biol.*, **1997**, *7*, 181.

[3] U. H. E Hansmann and Y. Okamoto, *Phys. Rev. E*, **1997**, *56*, 2228 – 2233.

[4] D. Frenkel and B. Smit, *Understanding Molecular Simulations;* Academic Press, San Diego, 2002.

[5] D. C. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, *Nature*, **2002**, *420*, 102–106.

[6] E. Paci, A. Cavalli, M. Vendruscolo, and A. Caflisch, *Proc. Natl. Acad. Sci. USA.*, **2003**, *100*, 8217–8222.

[7] G. Settanni, J. Gsponer, and A. Caflisch, *Biophys. J.*, **2004**, *86*, 1691–1701.

[8] A. R. Ferst, *Proc. Natl. Acad. Sci. USA.*, **2002**, *99*, 14122–14125.

[9] N. Singhal, C. D. Snow, and V. S. Pande, *J. Chem. Phys.*, **2004**, *121*, 415–425.

[10] W.C. Swope, J.W. Pitera, and F. Suits, *J. Phys. Chem. B*, **2004**, *108*, 6571–6581.

[11] S. Sriraman, I.G. Kevrekidis, and G. Hummer, *J. Phys. Chem. B*, **2005**, *109*, 6479–6484.

[12] J. D. Chodera, N. Singhal, V. S. Pande, K. Dill, and W. C. Swope, *J. Chem. Phys.*, **2007**, *126*, 155101.

[13] F. Noé, I. Horenko, C. Schuette, and Smith J.C., *J. Chem. Phys.*, **2007**, *126*, 155102.

[14] N.-V. Buchete and G. Hummer, *Phys. Rev. E*, **2007**, *77*, 030902.

[15] S. Muff and A. Caflisch, *Proteins: Structure, Function, and Bioinformatics*, **2008**, *70*, 1185–1195.

[16] Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, **1999**, *314*, 141–151.

[17] F. Rao and A. Caflisch, *J. Chem. Phys.*, **2003**, *119*, 4035–4042.

[18] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch, *J. Chem. Phys.*, **2004**, *121*, 10748–10756.

[19] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, *Proc. Natl. Acad. Sci. USA.*, **2005**, *102*, 6801–6806.

[20] D. van der Spoel and M. Marvin Seibert, *prl*, **2006**, *96*, 238102.

[21] S. Yang, J. N. Onuchic, A. E. Garcia, and H. Levine, *J. Mol. Biol.*, **2007**, *372*, 756–763.

[22] S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. USA.*, **2004**, *101*, 14766–14770.

[23] F. Rao and A. Caflisch, *J. Mol. Biol.*, **2004**, *342*, 299–306.

[24] A. Caflisch, *Curr. Opin. Struct. Biol.*, **2006**, *16*, 71–78.

[25] N.-V. Buchete and G. Hummer, *J. Phys. Chem. B*, **2008**, in press.

[26] M. Apaydin, D. Brutlag, C. Guesttin, D. Hsu, and J. Latombe, *"In International Conference on Computational Molecular Biology (RECOMB)"*, **2002**.

[27] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, *J. Phys. Chem. B*, **2008**, in press.

[28] S. V. Krivov and M. Karplus, *J. Phys. Chem. B*, **2006**, *110*, 12689–12698.

[29] P. Ferrara and A. Caflisch, *Proc. Natl. Acad. Sci. USA.*, **2000**, *97*, 10780–10785.

[30] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez, *Protein Science*, **1999**, *8*, 854–865.

[31] A. Cavalli, P. Ferrara, and A. Caflisch, *Proteins: Structure, Function, and Bioinformatics*,

2002, *47*, 305–314.

[32] A. Cavalli, U. Haberthür, E. Paci, and A. Caflisch, *Protein Science*, **2003**, *12*, 1801–1803.

[33] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch, *Bioinformatics*, **2007**, *(2007)*, in press.

[34] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*, **1983**, *4*, 187–217.

[35] E. Neria, S. Fischer, and M. Karplus, *J. Chem. Phys.*, **1996**, *105*, 1902–1921.

[36] P. Ferrara, J. Apostolakis, and A. Caflisch, *Proteins: Structure, Function, and Bioinformatics*, **2002**, *46*, 24–33.

[37] P. Ferrara, J. Apostolakis, and A. Caflisch, *J. Phys. Chem. B*, **2000**, *104*, 5000–5010.

[38] G. Settanni, F. Rao, and A. Caflisch, *Proc. Natl. Acad. Sci. USA.*, **2005**, *102*, 628–633.

[39] J. A. Ihalainen, B. Paoli, S. Muff, E. Backus, J. Bredenbeck, G. A. Woolley, A. Caflisch, and P. Hamm, *Proc. Natl. Acad. Sci. USA.*, **2008**, *105*, 9588–9593.

[40] J.A. Hartigan, *Wiley, New York*, **1975**.

[41] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA.*, **2006**, *103*, 17747–17752.

[42] S. F. Chekmarev, S. V. Krivov, and M. Karplus, *J. Phys. Chem. B*, **2005**, *109*, 5312–5330.

[43] C. A. F. Andersen, A. G. Palmer, S. Brunak, and B. Rost, *Structure*, **2002**, *10*, 174–184.

[44] Haberthür U and A. Caflisch, *J. Comput. Chem.*, **2008**, *29*, 701–715.

				Weight (%)		Mft (ns)	
	Sec. str. string	Name	NC[a]	REMD	CTMD	REMD	CTMD
330 K							
	-EEEESSEEEEEESSEEEE-	Native	yes	37.8	37.1	–	–
	-EEEESTTEEEEESSEEEE-	Ns-or$_1$	yes	1.9	2.2	115	106
	--EEESSSEEEEESSEEEE-	Ns-or$_2$	yes	3.6	2.7	126	109
	--EESSSEEEEESSEEEE-	Ns-or$_3$	yes	1.8	1.4	113	105
	-EEEESSEEEEESSEEEE-	Cs-or	yes	5.3	5.3	101	109
	---SSGGG---EESSEETT-	Ch-curl$_1$	no	2.5	0.6	175 (4.2%)[b]	263
	---SSGGG-EESSTTTTEE-	Ch-curl$_2$	no	1.4	1.2	N.A.[c]	201
286 K							
	-EEEESSEEEEEESSEEEE-	Native	yes	60.6		–	–
	-EEEESTTEEEEESSEEEE-	Ns-or$_1$	yes	3.1		705	2030
	--EEESSSEEEEESSEEEE-	Ns-or$_2$	no	2.8		6330 (98.0%)	3170
	--EESSSEEEEESSEEEE-	Ns-or$_3$	no	0.5		6370 (100%)	6690
	-EEE-STTEEEESSEEE--	Cs-or	no	0.7		13100 (96.4%)	970
	---SSGGG---EESSEETT-	Ch-curl$_1$	no	7.5		8820 (4.4%)	5260
	---SSGGG-EESSTTTTEE-	Ch-curl$_2$	no	4.0		N.A.[c]	7170

TABLE I: Comparison of populations and mft values from individual basins extracted from REMD simulations by ETN(A) and the corresponding values obtained by CTMD. The basins were identified with the cFEP approach and the DSSP secondary structure string [43] is the most frequent in the basin. [a]Several non-native basins at 330 K are in the native component (NC) of REMD, whereas the NC at 286 K consists of only the native basin and Ns-or$_1$. Note that the ETN or ETNA procedures were used for basins in the NC or outside of it, respectively. All folding times were extracted from the fit of the respective cumulative folding time distribution. [b]Values in parentheses are the fraction of snapshots to which a folding time could be assigned by ETNA. [c]The Ch-curl$_2$ basin was disconnected from the NC at all temperatures and therefore it is not possible to estimate its mft. Abbreviations: Ns-or, N-terminal strand out of register and folded C-terminal hairpin; Cs-or, C-terminal strand out of register and folded N-terminal hairpin; Nh-curl, curl-like conformation with folded N-terminal hairpin; Ch-curl, curl-like conformation with folded C-terminal hairpin.

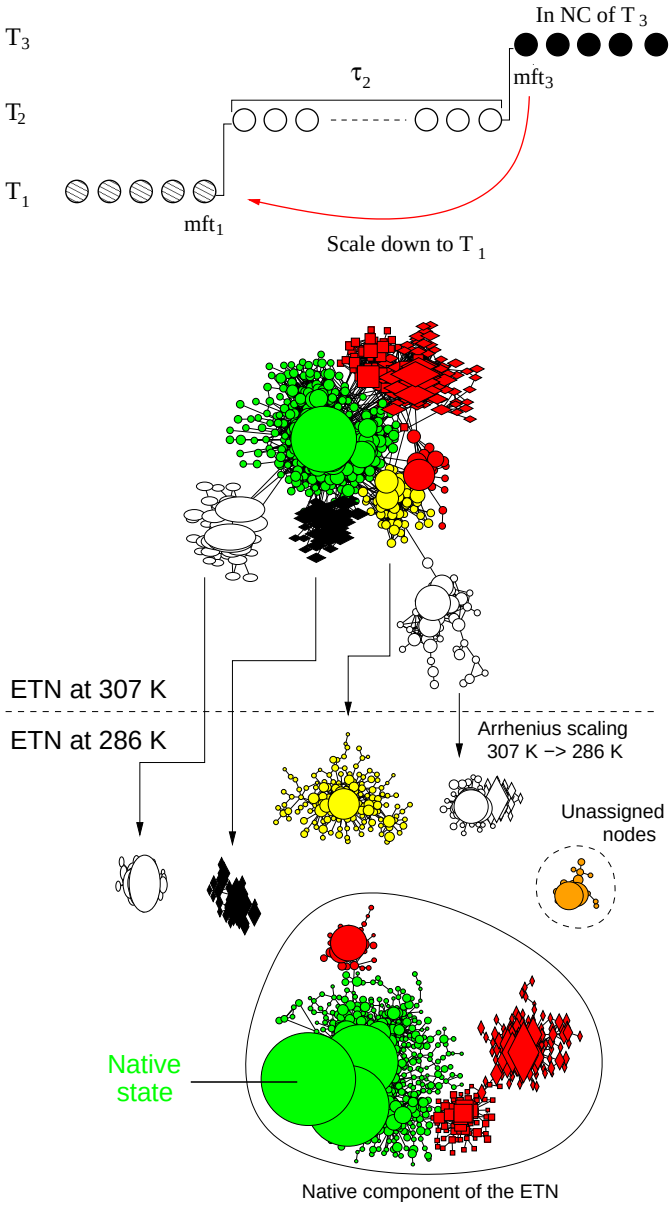


FIG. 1: Illustration of the ETNA procedure for folding time evaluation of snapshots outside of the NC at the temperature of interest. (Top) Whenever a disconnected REMD segment at temperature T_1 is visited (dashed nodes) the next segment in the time series of that replica is considered (white nodes). In the example shown, only the second temperature increase to T_3 is successful in visiting snapshots belonging to nodes of the NC (black nodes). The details of the procedure are given in the text. (Bottom) Schematic view of the main idea behind the Arrhenius scaling approach. Folding time information extracted at high temperature is used to estimate folding kinetics at low temperature. Each color and shape represents a free-energy basin.

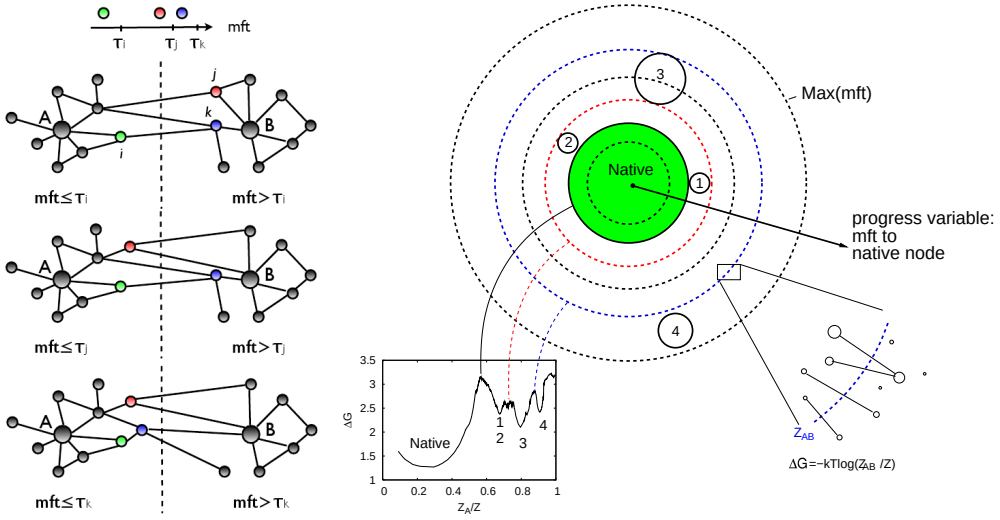


FIG. 2: Schematic illustration of the cFEP procedure [27, 28]. (Left) Nodes of the ETN are first sorted according to increasing mft. For each value mft_c between 0 (node A) and Max(mft) a value of the cut Z_{AB} between nodes A and B is calculated. The set of nodes on the left of the cut contains node A and all nodes with mft$\leq$mft$_c$, where Z_A/Z is its relative partition function. The green, red and blue nodes have consecutive values of mft in this simplified illustration of the ETN. (Right) Relation between free-energy basins and the cFEP. Each solid circle borders a basin, while concentric dashed circles represent values of mft. To illustrate the cFEP, $\Delta G = -kT log(Z_{AB}/Z)$ is plotted as a function of Z_A/Z. Basins 1 and 2 overlap because they have the same mft distance from the native state and are therefore not separated in the unfolded part of the profile.

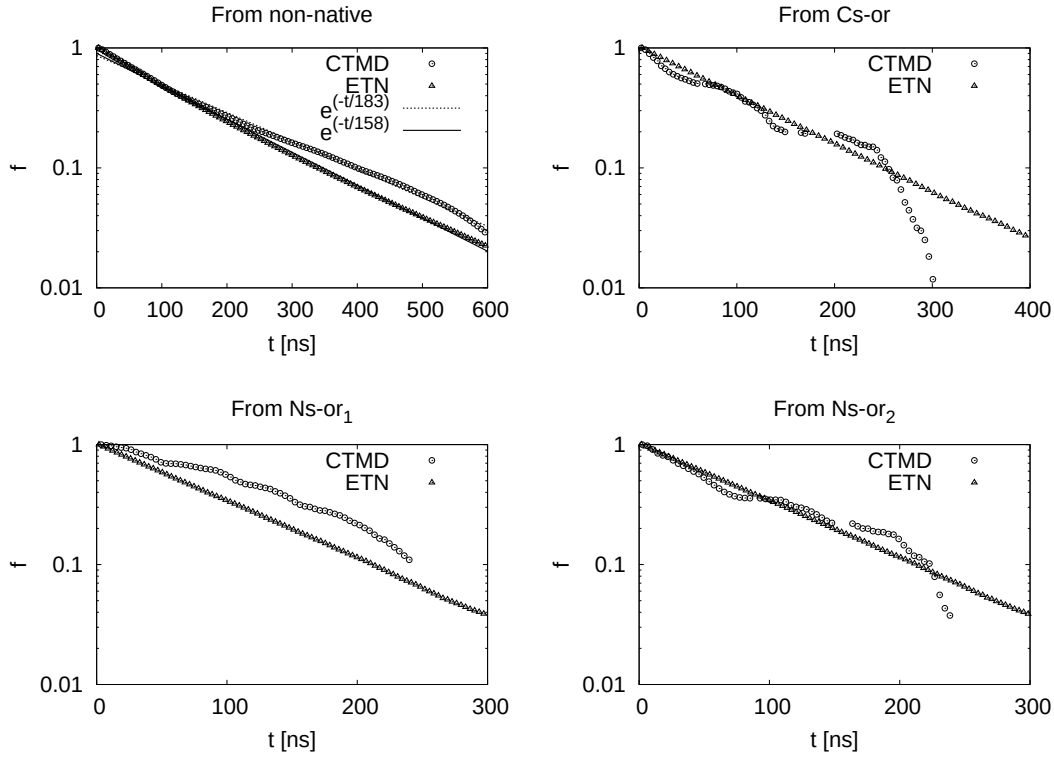


FIG. 3: Cumulative folding time distribution as extracted directly from the 330 K CTMD simulation (circles) and from the corresponding ETN, which is treated as a Markov state model (triangles). The folding dynamics from the non-native ensemble (top left) and from specific metastable states (top right and bottom) can be reproduced by the model, which is a very strong indication that the Markov assumption is justified for the lagtime of 20 ps used here.

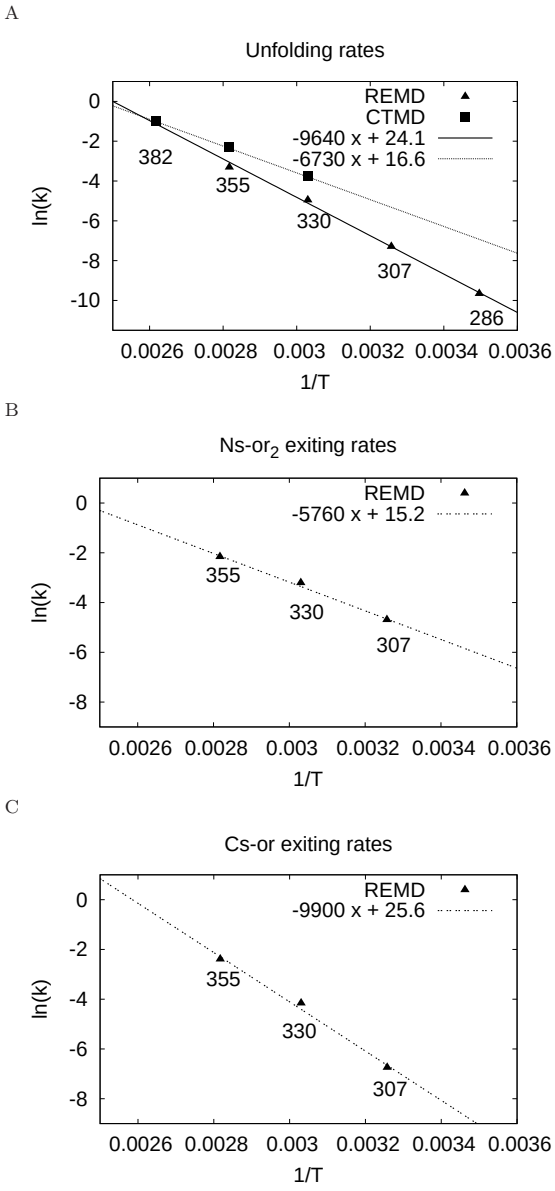


FIG. 4: Temperature dependence of rates to exit enthalpic basins. (A) Unfolding rates estimated by Monte Carlo runs on the ETN of the CTMD simulations (squares) and the ETN of the NC from individual REMD temperatures (triangles). The estimates for the Arrhenius constant E_a/R (Eq. (1)), which is used to scale the kinetics at different temperatures, can be extracted from the linear fit of unfolding rates. (B) Exiting rates from the Ns-or$_2$ basin. (C) Exiting rates from the Cs-or basin. Activation enthalpy values to exit individual basins are similar, justifying the use of only one Arrhenius constant in the ETNA approach.

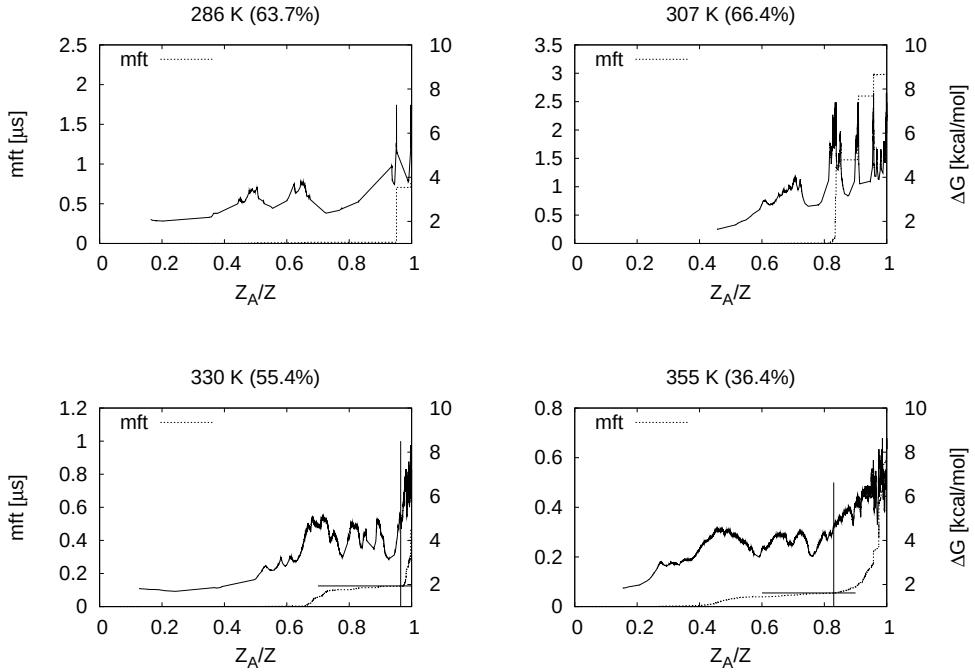


FIG. 5: Identification of enthalpic basins for Arrhenius scaling. At each value of the temperature, the cFEP of the NC is shown by solid lines with ΔG values on the right y-axis. The percentage values in parentheses represent the statistical weight of the respective NC. The plot of mft as a function of the relative partition function Z_A/Z is shown with dotted lines, and the selection of the enthalpically stabilized part of the ETN is indicated by perpendicular lines in black. The criterion was to include as many enthalpic minima as possible by cutting at the point (crossing of perpendicular lines) where the roughness of the cFEP indicates insufficient sampling, which is often the case in entropically stabilized regions. At 286 K and 307 K all nodes of the NC were included, whereas only a subset was used at 330 K and 355 K to remove entropic noise (see text for explanation). The mft cutoffs were chosen at 125 ns (330 K) and 55 ns (355 K). Interestingly, these cutoff values correspond roughly to the folding times observed in the CTMD simulations.

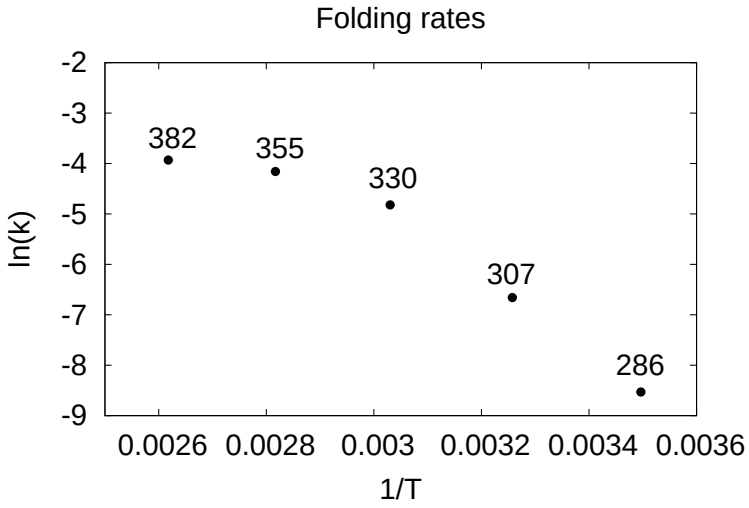


FIG. 6: Folding rates from CTMD equilibrium simulations (330, 355, 382 K) and folding runs (286, 307 K) calculated by exponential fitting of the cumulative folding time distribution. There is a clear anti-Arrhenius transition above 355 K, and therefore sampling at higher temperatures was not included in the ETNA analysis.

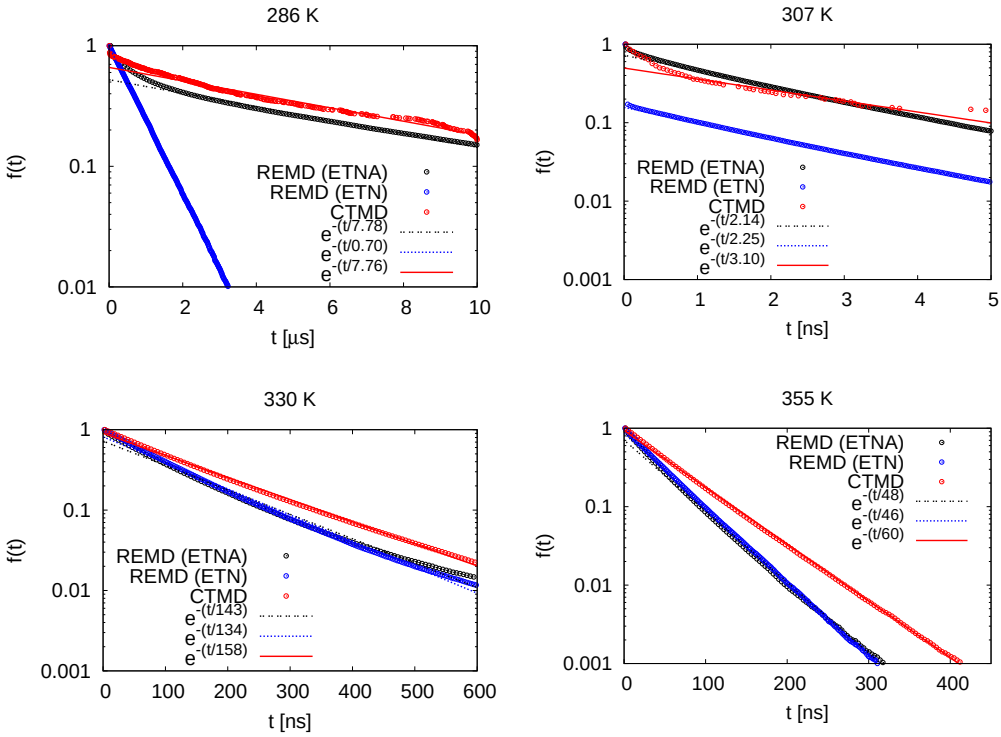


FIG. 7: The cumulative folding time distributions $f(t) = \int_t^\infty p(\tau)d\tau$ extracted from REMD using the ETNA approach (black) or only from the NC of the ETN (blue) are compared to the reference CTMD data (red). $p(\tau)$ is the probability density of the folding time distribution. The CTMD data at 286 K and 307 K were extracted from 750 and 250 folding simulations, respectively, started from the unfolded state ensemble. At 330 K and 355 K, equilibrium CTMD simulations of 20 μs and 10 μs, respectively, were performed to compare the folding time distribution to the REMD approach. The CTMD and ETNA curves at all temperatures are in remarkable agreement. The use of only the NC of the ETN at 286 K yields a folding time that is faster by a factor of ten, whereas for temperatures of 307 K or higher the use of the Arrhenius scaling (ETNA) and only the NC (ETN) are almost identical to the CTMD results.

ETNA: Equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations SUPPLEMENTARY MATERIAL

S. Muff* and A. Caflisch*

Department of Biochemistry, University of Zurich,

Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Keywords:

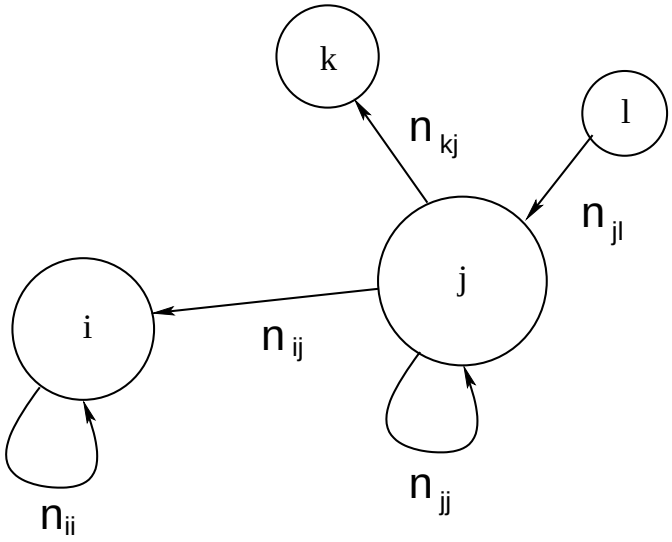


FIG. S1: Schematic representation of the transition matrix The absolute number of transitions from node j to node i is n_{ij}. The transition probability from node j to node i is calculated by $p_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$.

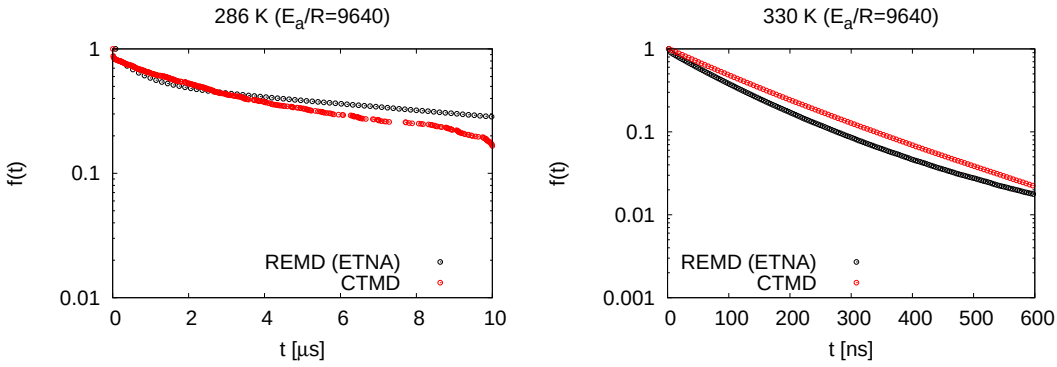


FIG. S2: Cumulative folding time distributions at 286 K (left) and 330 K (right) with ETNA using the Arrhenius parameter $E_a/R = 9640$ K, as calculated from the fit of unfolding rates on the ETNs of REMD.

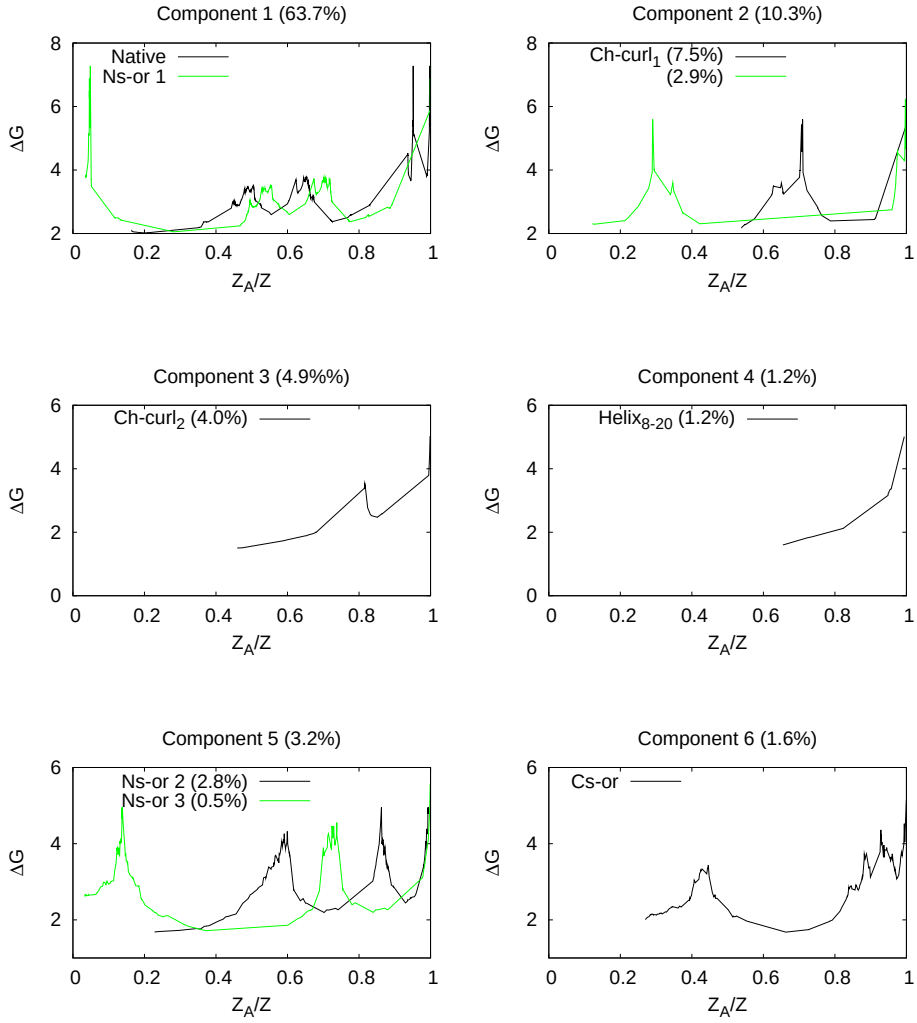


FIG. S3: cFEPs from the six most populated REMD components at 286 K. Component 1 is the NC.

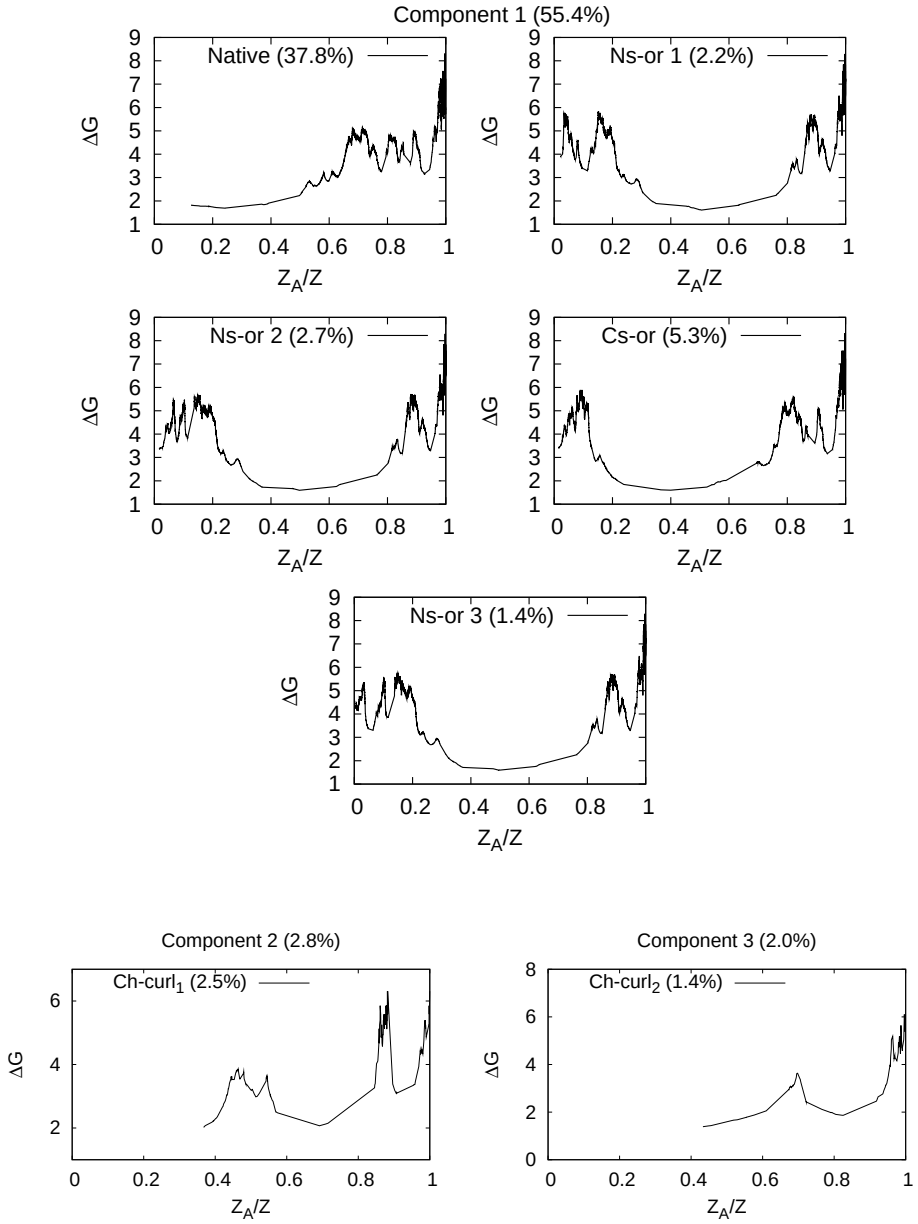


FIG. S4: cFEPs from the three most populated REMD components at 330 K. Component 1 is the NC.

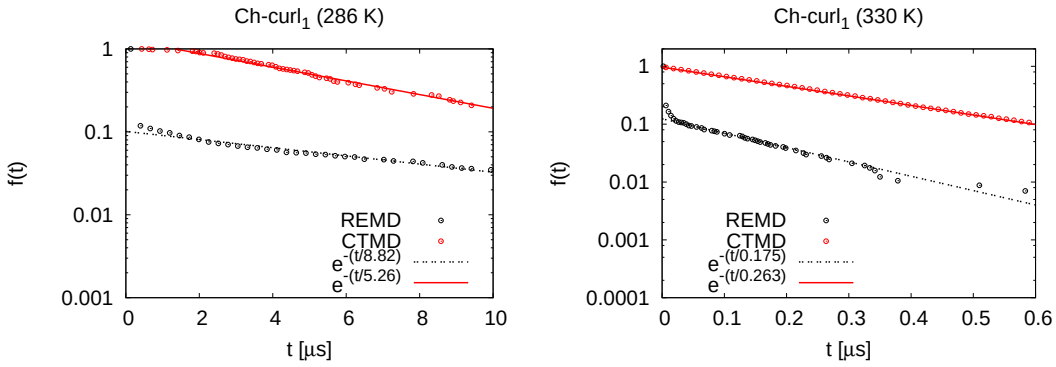


FIG. S5: Example fits to the cumulative folding time distributions from CTMD folding runs (red) and from ETNA scaling (black).

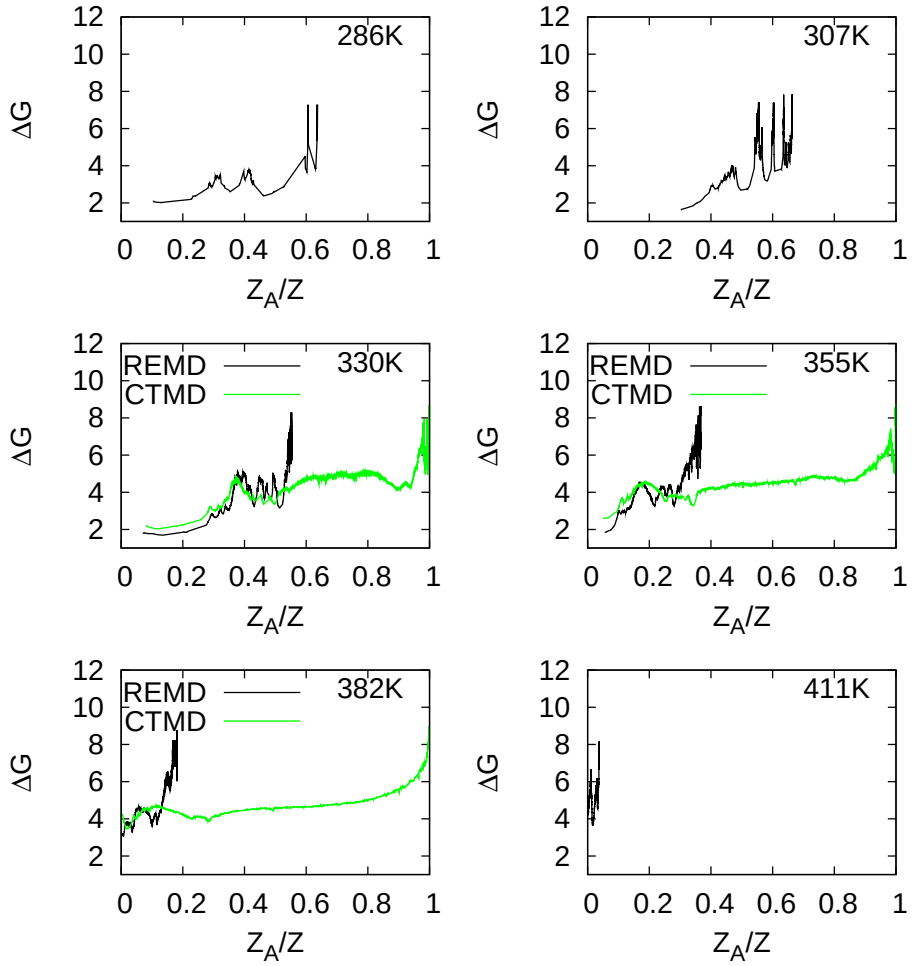


FIG. S6: cFEPs calculated from the REMD (black) and CTMD (green) simulation data. The REMD-cFEPs show only the NC, whose relative size is the Z_A/Z value of the rightmost black data point. At very low temperature (top) the ETN from REMD is disconnected because of high free-energy barriers. At high temperature (bottom), large entropic contributions have the same effect. All ΔG values are in kcal/mol.

* tel: +41 44 635 55 21 e-mail:caflisch@bioc.uzh.ch,smuff@bioc.uzh.ch

Chapter 7

Local modularity measure for network clusterizations.

[*Phys. Rev. E*, **2005,** *72*,056107]

Local modularity measure for network clusterizations

Stefanie Muff, Francesco Rao, and Amedeo Caflisch
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zuerich, Switzerland
(Received 10 March 2005; revised manuscript received 9 September 2005; published 7 November 2005)

Many complex networks have an underlying modular structure, i.e., structural subunits (communities or clusters) characterized by highly interconnected nodes. The modularity Q has been introduced as a measure to assess the quality of clusterizations. Q has a global view, while in many real-world networks clusters are linked mainly *locally* among each other (*local cluster connectivity*). Here we introduce a measure of localized modularity LQ, which reflects local cluster structure. Optimization of Q and LQ on the clusterization of two biological networks shows that the localized modularity identifies more cohesive clusters, yielding a complementary view of higher granularity.

DOI: 10.1103/PhysRevE.72.056107 PACS number(s): 89.75.Hc

Complex networks are a powerful tool for the analysis of a diverse range of systems, including technological [1,2], social [3,4], and biological networks [5,6]. Especially in biology, thanks to high-throughput experiments, there is a tremendous growth of available data that can be efficiently analyzed and summarized in terms of complex networks [7,8]. In many cases, networks have an inherent modular structure which can represent functional units called communities or clusters, e.g., web pages of a certain subject [9], social groups [3,10], or biological modules [11,12]. However, there is neither an obvious and commonly accepted definition of communities nor a straightforward way to find the underlying modules of a network. Recently, many clustering algorithms have been proposed [13–18]. For a clusterization with K communities, the *modularity* $Q=\sum_{i=1}^{K}[e_{ii}-(a_i)_{in}(a_i)_{out}]$ has been introduced as a measure to assess the quality of a clusterization [19], where $e_{ii}=L_i/L_{tot}$, the effective fraction of links inside community i, is compared to $(a_i)_{in}(a_i)_{out}=(L_i)_{in}(L_i)_{out}/L_{tot}^2$ which is the predicted fraction of edges that fall into community i if the links in a directed network are set between nodes without regard to the community structure. Q is high when the clusterization is good and it can reach a maximum value of 1. Modularity is used to compare the quality of different clusterizations, e.g., to find the best split of a dendogram [20] or to validate different clusterization methods and furthermore as a fitness function in optimization procedures, where Q_{max} should correspond to the objectively best clusterization of a network [11,14]. The modularity is a global measure because the comparison of L_i/L_{tot} with $(L_i)_{in}(L_i)_{out}/L_{tot}^2$ assumes that connections between all pairs of nodes are equally probable, which reflects connectivity among all clusters.

On the other hand, in many complex networks most clusters are connected to only a small fraction of the remaining clusters. In metabolic networks, for instance, major pathways occur as clusters that are sparsely linked among each other [11]. Furthermore, in the protein folding network [6] communities are energy basins and transitions, i.e., connections, are allowed only between adjacent basins [15]. We call this property *local cluster connectivity*. In this paper, we introduce a measure for the quality of network clusterizations. To take into account local cluster connectivity and to overcome global network dependency, the approach of modularity is modified into a *local* version. The contribution to modularity for each cluster i is calculated for the subnetwork consisting of cluster i and its neighbor clusters. This requires the determination of i's neighborhood or, more precisely, all the links L_{i_N} that are contained in this neighborhood. The sum of the contributions of all K clusters yields

$$LQ=\sum_{i=1}^{K}\left[\frac{L_i}{L_{i_N}}-\frac{(L_i)_{in}(L_i)_{out}}{(L_{i_N})^2}\right].$$

We call LQ *localized modularity*. It is – in contrast to Q – not bounded by 1, but can take any value. The more locally connected clusters a network has, the higher LQ is. On the other hand, in a network where all communities are linked among each other, Q and LQ coincide.

It is interesting to compare the behavior of Q and LQ on different network topologies and use them as fitness functions for the optimization of network clusterizations [11,14]. We start with an illustration of the differences between Q and LQ by discussing a simple example of a scalable local cluster connectivity network, which we call the *school network* [Fig. 1(a)]. It is a toy model of social interactions between pupils in a school with l levels and c classes per level. Levels have periodic boundary conditions to avoid spurious boundary effects (in the first and last levels). In a real school, all the students of a class know each other and, as a first approximation, a student would interact most with people of his or her age. In the school network model, students are the nodes of the network and a link between two pupils is made if they know each other. Each class contains s fully connected students. A link between two students of the same level but different classes is placed with a (high) probability $p\leqslant 1$ and connections between students that are one level above or below [+1, Fig. 1(a)] are made with smaller probability $r<p$. No social interaction is assumed between persons that are more than one level apart from each other, i.e., if one of the students is more than one year older than the other [+2 or more, Fig. 1(a)]. Interestingly, when only two levels and two classes per level are considered, the school network model is essentially the same as the well-known (globally connected) four communities test network used in [11,14]. Hence, the

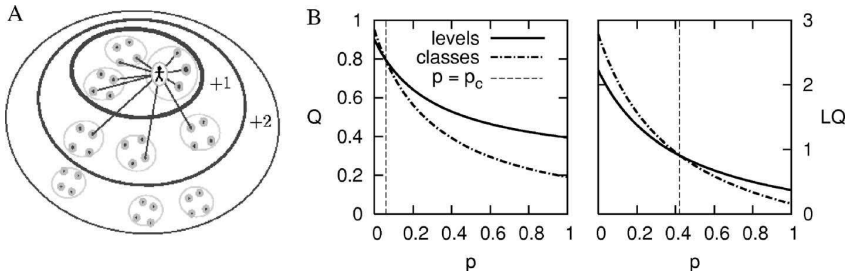


FIG. 1. (a) A student's view in the simplified schematic school network model with only three levels, three classes per level and four students per class: The student interacts with all his classmates, with other students on the same level with probability p=0.5, and with pupils one level above or below (+1) with probability r=0.25. No connections are assumed between students that are more than one level apart (+2 or more). (b) The p-dependent behavior of the modularity and the localized modularity in the school network with ten levels, two classes per level, 20 pupils per class, and $r=p/2$. The modularity favors the grouping of classes (solid line) in the same level for almost all p, whereas localized modularity favors communities consisting of single classes (dot-dashed line) for $p<0.42$.

school network is a simple generalization to locally connected networks. It is unweighted and undirected but an extension to directed and weighted networks, e.g., asymmetrical friendship, is straightforward.

A grouping of all the pupils on one level into the same cluster is reasonable for high p, i.e., when students of the same age interact among each other with high probability. But, as p decreases, classes become more and more separated from each other until they fully break apart for p=0, where a fitness measure is expected to favor clusterizations that identify classes. Therefore, we calculated modularity and localized modularity for the clusterization of nodes according to classes and according to levels for $p\in[0,1]$, $r=p/2$, and s=20 students per class. Figure 1(b) shows the Q and LQ values for ten levels and two classes per level. They were obtained analytically, using the expected numbers of links for each p. Both Q and LQ favor the clusterization into levels for p close to 1. LQ yields the same value for both clusterizations (*crossing point*) at p_c^{LQ}=0.42 and prefers the clusterization into classes for $p<0.42$. The modularity, on the other hand, has its crossing point at p_c^{Q}=0.09, i.e., it favors the classes only for $p<0.09$. In other words, Q considers the classes and not the levels as the best cluster partition only if the probability of interaction between two students of the same age but different classes is smaller than 10%.

The crossing point p_c depends on the number of levels and classes. Figure 2 shows the change of p_c upon variation of these two parameters with two, five, and ten classes per level, respectively (from top to bottom). It can be seen that p_c^{LQ} is higher than p_c^{Q} for all values of levels and classes, and is by construction constant for a fixed number of classes per level. On the other hand, p_c^{Q} strongly depends on network size which means that it favors different clusterizations as the number of levels increases, i.e., the lens of cluster detection becomes more coarse. Furthermore, it converges to 0 as l grows, meaning that Q favors the clusterization into levels for any $p\in[0,1]$, even though the classes on the same level are almost disconnected for small p.

These observations indicate that LQ is more reliable than Q to validate clusterizations in local cluster connectivity networks. The discrepancies between the two measures originate from the fact that Q compares the effective to the expected fraction of links in the clusters, no matter if a link is *possible* or not. The expected fraction of links is therefore underestimated in local cluster connectivity networks, thus the difference between the expected and the effective fraction of links (i.e., Q) is overestimated. On the other hand, LQ only takes into account local link expectations. Furthermore, note that modularity as high as 0.8 has been found in Erdös-Réni (ER) random graphs, scale-free networks, and regular lattices [21,22].

In recent years, biological networks [23] have attracted the attention of many scientists for their potential impact on the understanding of living systems. Metabolic and protein-protein interaction networks have been clustered by Q optimization [11] and the MCL method [24], respectively. To investigate the behavior of Q and LQ on real-world networks we optimized the clusterizations of two recent realizations of the metabolic and protein-protein interaction networks of E. coli by simulated annealing (SA), using each of the two measures as cost function. For each temperature T, c_1n^2 single-node and c_2n multinode moves, like splitting and merging of (adjacent) communities, were performed, where $c_{1,2}$ are constants and n is the number of nodes in the network. Furthermore, T was iteratively reduced to c_3T with a constant

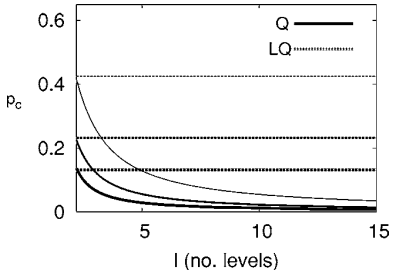


FIG. 2. Dependence of p_c on network size: for two, five, and ten classes per level (from top to bottom), p_c^{LQ} (dotted lines) is always higher than p_c^{Q} (solid lines) showing that LQ favors the clusterization into classes for higher p while Q almost always prefers the grouping into levels. Moreover, p_c^{Q} is rather sensitive on the size of the network and converges to 0 as the network grows, while p_c^{LQ} does not depend on the number of levels.

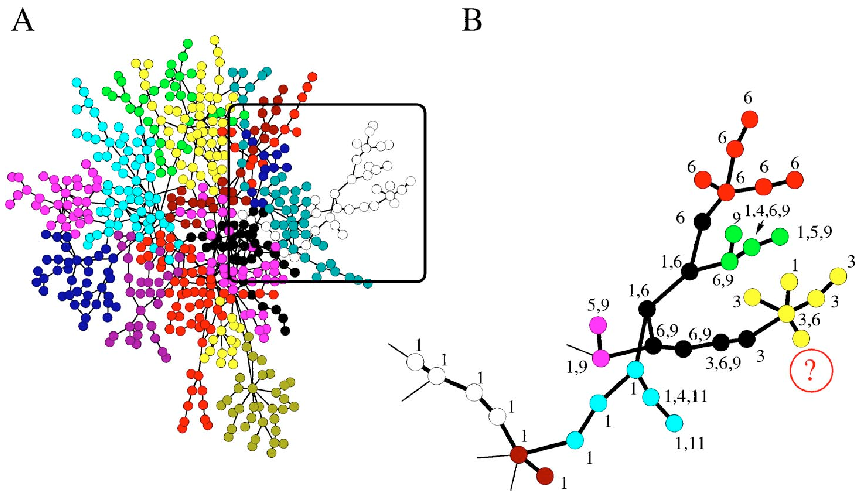


FIG. 3. (Color online) (a) Largest connected component of the metabolic network of *E. coli*. The coloring scheme represents the clusterization found by optimizing modularity. Some colors are used twice. (b) *LQ* clusterization of the white *Q* cluster with the annotation of different pathways. According to *LQ* it is highly probable that the unassigned yellow node (*N*-acetyl-alpha-*D*-glucosamine 1-phosphate, marked as "?") belongs to the carbohydrate metabolism (label 3).

$c_3 < 1$. This move set and cooling scheme is similar to the one used in [11]. The computational effort for the two measures scales as $O(K)$, even though the calculation of *LQ* is slightly more expensive since it involves the determination of neighborhoods for each cluster.

(i) *The metabolic network of E. coli.* We use the metabolic pathway database developed by Ma and Zeng [25], which has been derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [26]. Figure 3 shows the largest connected component of the *E. coli* metabolic network in this database. It contains 563 nodes and 708 links which have been treated undirected. Each node is assigned to between zero and nine out of 11 possible pathways. The optimization with fitness function *Q* leads to a division into 16 clusters consisting of 35 metabolites on average (as colored in Fig. 3) and takes a value as high as $Q_{max}=0.82$. On the other hand, *LQ* optimization leads to a maximum of $LQ_{max}=12.1$ with 132 clusters, each containing an average of 4.3 metabolites. The optimization of the two measures finds clusters at a different level, which yields complementary information. As expected, *Q* is based on a global view and depends on the size of the network. As a consequence, optimizing a network with more metabolites would lead to larger *Q* clusters. This problem is likely to arise because, as more data become available, the network and its largest connected component will grow. On the other hand, *LQ* finds the lowest-level modules, independent on the rest of the network. Still, a major motivation to find clusters is to obtain information about presumed pathways of nonannotated metabolites. Figure 3(b) zooms into one of the *Q* clusters (white) and shows the splitting into smaller *LQ* clusters. The numbers indicate the respective pathway(s) of the nodes. Note that an *LQ* cluster is not necessarily fully contained in a *Q* cluster, i.e., a smaller (local) cluster may be only *partially* contained in a larger one. In the considered cluster of Fig. 3(b), the further division is justified because it results in more homogeneous subclusters. The yellow community, for instance, contains mainly nodes belonging to the carbohydrate metabolism pathway (label 3). According to this, the unassigned node [*N*-acetyl-alpha-*D*-glucosamine 1-phosphate, labeled as "?" in Fig. 3(b)] can also be classified in pathway 3 with high confidence. This would have been impossible when considering the white cluster obtained by *Q* whose nodes are assigned mainly to pathway 6 (glycan biosynthesis and metabolism) and 1 (amino-acid metabolism).

To obtain a more quantitative analysis, we compute the conditioned probability

$$P[i,j] = P[\pi(i) \cap \pi(j) \neq \emptyset \,|\, c(i) = c(j)] \qquad (1)$$

that two nodes *i* and *j*, lying in the same cluster *c*, share at least one pathway (π). For the *Q* clusterization, this probability is $P_Q[i,j]=0.57$, while $P_{LQ}[i,j]=0.73$, reflecting the higher homogeneity of the *LQ* clusters. Comparison to the null case, where nodes are picked at random from the network, yields $P_R[i,j]=0.26$ and the probability that any pair of linked nodes shares a pathway is 0.59, thus essentially the same as for the clustering with *Q*.

(ii) *The protein-protein interaction (PPI) network of E. coli.* A set of 716 verified interactions involving 270 proteins of *E. coli* has been reported [27]. We again focused on the largest connected component consisting of 230 proteins and 695 undirected connections (Fig. 4). Identifying clusters can help to find indications about the function of unknown proteins. Again, modularity and localized modularity differ in the granularity of the clusters, similar to using two different lenses of a microscope. While the highest value for *Q* has been found for a clusterization with seven communities ($Q_{max}=0.49$), *LQ* splits the network into 56 communities ($LQ_{max}=2.97$). An example where *LQ* yields a more accurate "guess" is given in Fig. 4(b), where the *LQ* clusterization further subdivides the black cluster of Fig. 4(a). The proteins in the green circle are part of the DNA polymerase complex (dnaE, dnaQ, dnaX, dnaQ, holA, holB, holC, holD and holE). According to *LQ*, the unknown protein b1808 appears to be a protein of this complex. On the other hand, the black cluster obtained by *Q* is more heterogeneous which makes a functional assignment of b1808 difficult.

In conclusion, a measure for the quality of network clusterizations, called *localized modularity*, has been introduced and compared to the widely used *modularity*. Both measures can be used essentially in the same way. The latter has been

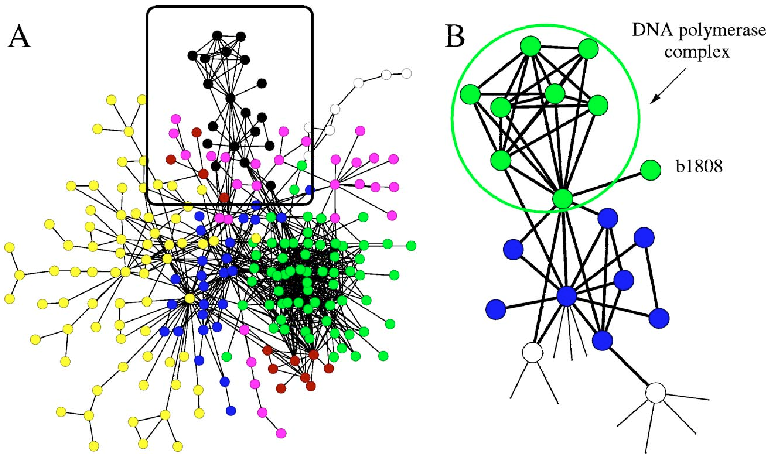


FIG. 4. (Color online) (a) Largest connected component of the PPI of *E. coli*. The colors represent the clusterization found by optimizing modularity. (b) *LQ* clusterization of the black *Q* cluster. The green circle contains proteins belonging to the DNA polymerase complex. The unknown protein b1808 is assigned to this complex according to *LQ* while the complete *Q* cluster is heterogeneous.

applied previously by others to assess the clusterization quality in many networks and has been used to find the best split of a dendogram and as fitness function in optimization algorithms. Finding clusters by optimizing a given fitness function has the advantage of not using any parameters (unlike many other clustering methods [15,17,18]). *Q* depends on global properties like the network size and the cluster connectivity. However, in many real-world networks, communities are merely connected locally, i.e., most pairs of clusters are not linked. We have called such organization *local cluster connectivity*. By detailed investigation of model networks as well as the optimization of *Q* and *LQ* on two biological networks, we have provided evidence that the two measures give a view of different depth into the cluster structure. In contrast to *Q*, *LQ* takes into account individual clusters and their nearest neighbors, generating high-confident clusters, irrespective of the rest of the network. Thus, the two measures provide complementary information. Furthermore, the *LQ* approach can be generalized to second or higher nearest neighbors which, albeit computationally more expensive, might yield additional insights, as if one were to use different lenses of a microscope.

This work was supported by a grant from the Swiss National Science Foundation.

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (2004).
[2] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[3] J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, London, 2000).
[4] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
[5] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **427**, 839 (2004).
[6] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[7] Y. Xia, H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, H. Zhao, and M. Gerstein, Annu. Rev. Biochem. **73**, 1051 (2004).
[8] A.-L. Barabási and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[9] J.-P. Eckmann and E. Moses, Proc. Natl. Acad. Sci. U.S.A. **99**, 5825 (2002).
[10] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, UK, 1994).
[11] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[12] M. B. Eisen, P. T. Spellman, P. O. Brow, and D. Botstein, Proc. Natl. Acad. Sci. U.S.A. **95**, 14863 (1998).
[13] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[14] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[15] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[17] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, Nucleic Acids Res. **30**, 1575 (2002).
[18] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2004).
[19] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[20] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[21] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).
[22] C. P. Massen and J. P. K. Doye, Phys. Rev. E **71**, 046101 (2005).
[23] M. G. Grigorov, Drug Discovery Today **10**, 365 (2005).
[24] J. P. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, Proteins: Struct., Funct., Bioinf. **54**, 49 (2004).
[25] H. Ma and A.-P. Zeng, Bioinformatics **19**, 270 (2003).
[26] M. Kanehisa and S. Goto, Nucleic Acids Res. **28**, 27 (2000).
[27] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, Nature (London) **433**, 531 (2005).

Conclusions and Outlook

The focus of the work in this thesis was directed to the development and improvement of methods that help to understand the very complex behavior of systems simulated at atomistic resolution. The huge number of degrees of freedom complicates their analysis and it was shown many times that arbitrarily selected progress variables are insufficient for describing the free-energy surface appropriately [1–4].

One main conclusion of this thesis is that free-energy surfaces should not be analyzed according to geometrical characteristics, but according to (local) equilibrium transitions. A new method, the *kinetic grouping analysis* (KGA) [5] was presented. KGA groups nodes (i.e., coarse-grained conformations) according to fast relaxation at equilibrium. The method was successful in the identification of metastable states from an equilibrium simulation of the three-stranded β-sheet peptide Beta3s [5] and from MD folding runs of an α-helical peptide [6].

Another method discussed and further developed is based on *cut-based free-energy profiles* (cFEPs) [7, 8]. The cFEP method was shown to fully quantify the free-energy surface, including basins, barriers and the transition state of folding. Furthermore, the cFEP method was applied to the description of constant temperature data extracted from *replica exchange molecular dynamics* (REMD) simulations. The REMD method enhances the sampling of the conformational space at low (physiological) temperature and reproduces correct thermodynamics, but not the kinetics, although there have been some efforts in this direction [9–11]. One chapter of the thesis addressed the problem of kinetics extraction from REMD by combining the equilibrium transition network (ETN) with an Arrhenius-based treatment in order to scale high-temperature kinetics to lower temperature. The approach was called ETNA.

The long-term aim is to apply the presented methods to the study of larger and more realistic systems, in order to obtain insight into biologically relevant processes of proteins. The root mean square deviation (rmsd) clustering applied in some of the presented studies, however, is limited to small systems, because by definition the rmsd is an average over all atoms included

in the calculation. Therefore, interesting structural changes are likely to be averaged out. A possible solution is to reduce the number of atoms by focussing only on the relevant subpart of the system that is involved in the conformational change. This last approach was already successfully applied to MD simulations of the aspartic protease β-secretase (BACE), where only dynamical changes in the flap region were of interest. A similar analysis is about to be applied to a large conformational transition of the B-Raf kinase, where only a few residues, which are involved in the rearrangement of a small loop, are considered.

Bibliography

[1] F. Rao and A. Caflisch. The protein folding network. *J. Mol. Biol.*, 342:299–306, 2004.

[2] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA.*, 101:14766–14770, 2004.

[3] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch. Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.*, 122:184901, 2005.

[4] A. Caflisch. Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.*, 16:71–78, 2006.

[5] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics*, 70:1185–1195, 2008.

[6] J. A. Ihalainen, B. Paoli, S. Muff, E.H. Backus, J. Bredenbeck, G.A. Woolley, A. Caflisch, and P. Hamm. Alpha-helix folding in the presence of structural constraints. *Proc. Natl. Acad. Sci. USA.*, 105:9588–9593, 2008.

[7] S. V. Krivov and M. Karplus. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B*, 110:12689–12698, 2006.

[8] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-dimensional barrier preserving free-energy projections of a beta-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B*, 112:8701–8714, 2008.

[9] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. USA.*, 102:6801–6806, 2005.

[10] S. Yang, J. N. Onuchic, A. E. Garcia, and H. Levine. Folding time predictions from all-atom replica exchange simulations. *J. Mol. Biol.*, 372:756–763, 2007.

[11] N.-V. Buchete and G. Hummer. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, 112:6057–6069, 2008.

Acknowledgments

This work was only possible thanks to the fantastic support by Prof. Amedeo Caflisch, who put a lot of confidence in my work since the very beginning. I also thank Prof. Ben Schuler, for being part of my PhD committee and contributing interesting thoughts during my seminars.

Many people of the group have added a piece to the mosaic of this thesis. Special thanks go to Francesco Rao, who guided me through my first year. Great computer support was given by Urs Haberthür, Francesco Rao, Fabian Dey and Philipp Schütz. Michele Seeber is acknowledged for providing the analysis program WORDOM and helping me to implement new functions. I thank François Marchand for being patient sharing the office with me during four years. All the rest of the group provided a very friendly and helpful environment.

I thank Sergei Krivov and Martin Karplus for a very kind and exciting collaboration.

Special thanks go to my parents, Frieda and Hans Muff, who supported me during all my life, and my partner Emil.

Curriculum Vitae

MUFF Stefanie

born 5^{th} November 1978 in Schwyz, Switzerland
Swiss citizen

Education

May 2004–now	Ph.D. thesis at the Department of Biochemistry, University of Zurich.
April 2008	Final examination for the Diploma of "Höheres Lehramt" (teaching diploma) in Mathematics
October 2003	Final diploma examination, passed with highest honors
1998–2003	Studies of Mathematics at the University of Zurich
1993–1998	Kantonsschule Kollegium Schwyz, specializing in natural sciences

Appendix A

WORDOM manual (KGA and cFEPs)

A.1 Kinetic Grouping Analysis (KGA)

KGA can be used for the identification of free-energy basins, not according to geometrical characteristics (such as the fraction of native contacts or RMSD from the folded structure) but rather according to fast relaxation at equilibrium. More explicitly, two coarse-grained conformations are grouped if along the MD trajectory their snapshots interconvert in more than 50% of the cases within a short commitment time τ_{commit}, which represents a typical relaxation time within basins of the investigated system [1]. The idea behind this approach is that if two conformations interconvert rapidly, they are not separated by a barrier and therefore belong to the same basin.

A bit of Theory

The commitment time τ_{commit}. The typical relaxation time within basins mentioned above, τ_{commit}, is a characteristic of the investigated system. It is an important parameter of KGA and defines the lens of resolution with which basins are isolated. A short τ_{commit} will group structures only locally or if the free-energy surface is very smooth. A longer τ_{commit} is more generous and might group subbasins separated by (low) barriers into larger basins. The first passage time to the native node (or a representative node of another basin), plotted as a free energy on a logarithmic x-axis, usually reflects two timescales: the inter- and intrabasin relaxation times (see Figure A.1). The barrier separating the two regimes can give a good indication for an (upper bound) of a typical relaxation time.

Isolation of all relevant basins at once. For a fixed commitment time τ_{commit} a matrix with interconversion (commitment) probabilities p_{commit}

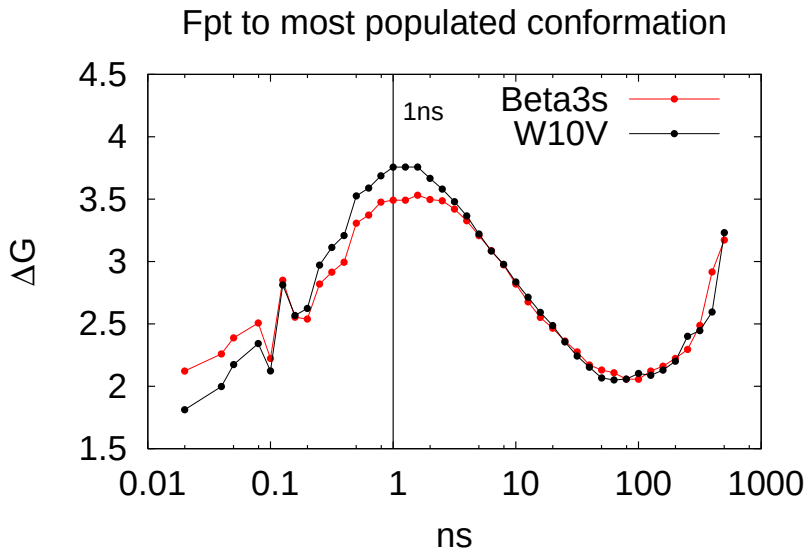


Figure A.1:

between any pair of nodes can be calculated, and pairs of nodes with $p_{commit} >= 0.5$ are grouped together. The grouping is a transitive procedure, i.e., if $p_{commit}(i,j) >= 0.5$ and $p_{commit}(j,k) >= 0.5$, then are also i and k in the same basin, even if $p_{commit}(i,k) < 0.5$. Since the computational cost of all-against-all calculations increases quadratically, in practice one selects a subset of highly populated nodes (e.g., the 500 most populated nodes), calculates the p_{commit}-matrix and divides them into basins. In a post-processing step, all other nodes are assigned commitment probabilities to the isolated basins and grouped to a basin if $p_{commit} >= 0.5$ for one of the basins. Otherwise, these nodes remain unassigned. Both the heavy-node calculation and the post-processing is done by WORDOM in the same function.

Isolation of a single basin. If only one basin is of interest or the basins have different relaxation times and one wants to isolate them one-by-one, p_{commit} is calculated only with respect to a given node (typically the representative/most populated node of a basin) and all nodes with $p_{commit} >=0.5$ are grouped into the investigated basin.

How to use the module

First passage time-plot to find the commitment time

Options:

```
-logbin     timeseries of noderanks
-bpd        bins per decade
-target     nodename with respect to which the first passage
            time should be calculated
```

Reads in the timeseries of noderanks (i.e., most populated node=1, second most populated node=2, etc.) and gives back the x- and y-coordinates of the logarithmically binned free-energy fpt-plot with respect to a selected node. ATTENTION: If pieces of trajectories from different simulations are concatenated, insert a line with the entry "0" to prevent them from being used as a continues timeseries. "0" must not be used otherwise.

Example:

```
wordom -logbin noderank.tt -bpd 10 -target 1
```

Kinetic grouping analysis to isolate all basins at once

Options:

```
-ka         timeseries of noderanks
-tcomm      commitment time (number of frames)
-nnodes     number of nodes for all-against-all
```

Reads in the timeseries of noderanks and groups nodes into basins according to KGA. The procedure calculates the all-against-all matrix for a selected number of most populated nodes and assigns all other nodes in a post-processing step. The output file contains first the nnodes x nnodes matrix of commitment probabilities. In the second part all nodes with their respective basin number are printed.
ATTENTION: If pieces of trajectories from different simulations are concatenated, insert a line with the entry "0" to prevent them from being used as a continues timeseries. "0" must not be used otherwise.

Example:

```
wordom -ka noderank.tt -tcomm 50 -nnodes 500
```

Kinetic grouping analysis to isolate a single basin

Options:

```
-ka1       timeseries of noderanks
-tcomm     commitment time (number of frames)
-target    nodename (rank) of the target node
```

Reads in the timeseries of noderanks. The output is the list of commitment probabilities (p_{commit}) of all nodes to the target node. The last part of the output is a list of all nodes in the basin of the selected target node.
ATTENTION: If pieces of trajectories from different simulations are concatenated, insert a line with the entry "0" to prevent them from being used as a continues timeseries. "0" must not be used otherwise.

Example:

```
wordom -ka1 noderank.tt -tcomm 50 -target 1
```

A.2 Cut-based free-energy profiles (cFEPs)

A progress coordinate that preserves the barriers and minima in the order that they are met during folding/unfolding events was introduced by Krivov and Karplus. It uses the relative partition function as the progress coordinate and determines the free energy barriers as a function of the coordinate by a method based on p_{fold}. The procedure gives almost identical results if p_{fold} is replaced by the mean first passage time (mfpt) to a selected node [2].

Pfoldf (p_{fold} fast)

Given the transition network with symmetrized links (equilibrium kinetic network or EKN) and two nodes A and B, corresponding to the "folded" and "denatured" node, the p_{fold} of node i is the solution of the equation $p_i = \sum_j p_{ji} \cdot p_j$ with boundary conditions p_A=1 and p_B=0. In a 2-state system with two enthalpic basins, one corresponding to the folded and one to the unfolded state, the two nodes A and B are the representative (most populated) nodes of the system.

However, in many systems, a node such as B does not exist, because there are multiple basins and/or an entropic state that cannot be represented by

a single node. Thus, as in the balanced minimum-cut procedure [3], an extra node B is introduced and connected to all nodes in the network with capacity $\lambda\tilde{w}$, where λ is a Lagrange multiplier (usually<0.01). The p_{fold} calculations are performed on the EKN with the extra node and the nodes are sorted according to their p_{fold}. Each value p_c between 0 and 1 can then be used to cut the network into set A containing all nodes with $p_{fold} \geq p_c$ and set B containing the nodes with $p_{fold} < p_c$. For each cut a point (x=Z_A/Z, y=-$kTln(Z_{AB}/Z)$) of the cFEP is obtained; Z_A/Z is used as the progress coordinate and Z_{AB} is the number of EKN-transitions between the two sets. The minimal cut value Z_{AB} between two sets split by the p_{fold} variable is a good approximation of the minimal cut between A and B [2,4], implying that the maximal value of $-kTln(Z_{AB}/Z)$ is a good approximation of the barrier.

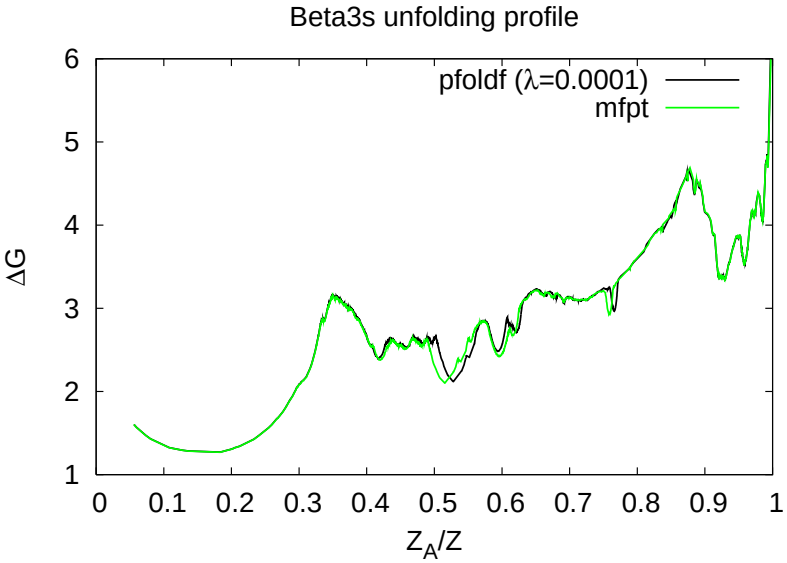


Figure A.2:

Mfpt (Mean first passage time)

In the pfoldf procedure described above the input are two nodes. Pfoldf is therefore appropriate to find barriers between two well-defined basins. However, sometimes it is useful to plot unfolding profiles with respect to only one node, especially if no representative node in the denatured state exists. Pfoldf solves the problem by introducing the extra node, but it is simpler to change the progress variable and use the mean first passage time (mfpt) to the node of interest, because mfpt is defined only with respect to one node.

Given the EKN, the mfpt of node i is the solution of the equation $\text{mfpt}_i = \Delta t + \sum_j p_{ji} \cdot \text{mfpt}_j$ with boundary condition $\text{mfpt}_A = 0$ [5]. The timestep Δt corresponds to the saving frequency of 20 ps, i.e., the mfpt of a node is defined as one timestep plus the weighted average of the mfpt values of its adjacent nodes. Mfpt has explicit time dependence through the occurrence of Δt in the equations. The resulting large system of linear equations differs from the one of pfoldf only by the Δt constant and the boundary conditions. Therefore, both can be solved with the same efficiency by iterative multiplication. Mfpt does not require an extra node, because mfpt is not defined between a pair of nodes, but only with respect to one node. To calculate the cFEP, the nodes are sorted according to their mfpt value. For all node-values mfptc a point (x=Z_A/Z, y=-$kTln(Z_{AB}/Z)$) on the cFEP can be calculated, where A is the set of all nodes with $\text{mfpt}_i \leq \text{mfpt}_c$ and B the set of nodes with $\text{mfpt}_i > \text{mfpt}_c$. The differences between unfolding cFEPs of the Beta3s peptide for pfoldf with λ=0.0001 and mfpt are marginal (see Figure A.2).

The mfpt-mfpt plot. Mfpt is the progress variable of the mfpt procedure, while the progress coordinate is the normalized partition function Z_A/Z. It is also possible to substitute the Z_A/Z coordinate by the more informative mfpt-values of the nodes by performing a transformation $x \rightarrow mfpt(x(Z_A/Z))$ where x(Z_A/Z) assigns nodes to each position on Z_A/Z, i.e., the mfpt that was originally used to rank on the Z_A/Z axis is now directly assigned to the nodes. An example of this is given in Figure A.3.

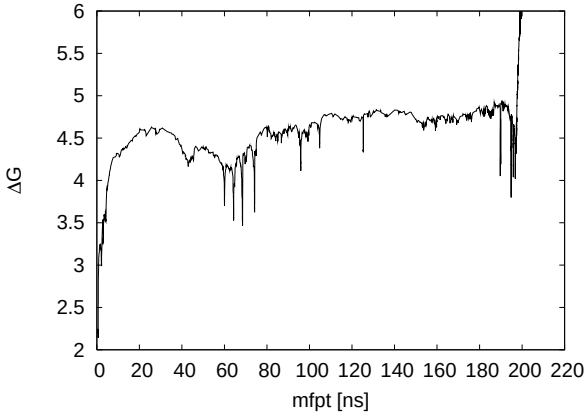


Figure A.3:

How to use the module

Pfoldf

Options:

```
-pfoldf     timeseries of noderanks or linkfile
-fepnet     necessary option, if the linkfile is given as input,
            otherwise not
-nonsymm    prevents symmetrization of network, i.e., detailed balance
            is not applied (default uses symmetrized network)
-target     start node (pfold=1)
-target2    stop node (node with pfold=0; default: 0=extra node)
-lambda     Lagrange multiplier
            (default value: if target2=0: l=0.0001; else l=0)
-nit        number of iterations to solve the equations
            (default value: 50'000)
-temp       temperature of the system (default value: 300K)
```

Examples:

```
wordom -pfoldf noderank.tt -target 1 -target2 0 -lambda 0.0001 -nit 100000
wordom -pfoldf linkfile -fepnet -nonsymm -target 1 -target2 0 -temp 330
```

Note that the output file contains nodes sorted according to their weight (number of snapshots). Therefore, to plot the profile it is possible to stop the calculation after a desired number of output pairs and then sort according to column 1 (e.g., — head -2000 — sort -nk1). Four rows are printed: \$1=Z_A/Z; \$2=ΔG; \$3=p_{fold}; \$4=node name.
ATTENTION: If pieces of trajectories from different simulations are concatenated, insert a line with the entry "0" to prevent them from being used as a continues timeseries. "0" must not be used otherwise.

Mfpt

Options:

```
-mfpt       timeseries of noderanks or linkfile
-fepnet     necessary option, if the linkfile is given as input, otherwise not
-nonsymm    prevents symmetrization of network, i.e., detailed balance
            is not applied (default uses symmetrized network)
-target     reference node of the mfpt procedure
-nit        number of iterations to solve the equations (default value: 50000)
-temp       temperature of the system (default value: 300K)
```

Examples:

```
wordom -mfpt noderank.tt -target 1 -nit 100000 -temp 330
wordom -mfpt linkfile -fepnet -nonsymm -target 1 -nit 100000 -temp 330
```

As for pfoldf, the output file contains nodes sorted according to their weight (number of snapshots). Therefore, to plot the profile it is possible to stop the calculation after a desired number of output pairs and then sort according to column 1 (e.g., — head -2000 — sort -nk1). To use mfpt as reaction coordinate instead of Z_A/Z: the columns in the output file are \$1=Z_A/Z, \$2=ΔG, \$3=mfpt. Therefore, (x=\$1,y=\$2) is the usual ΔG vs. Z_A/Z plot, while (x=\$3, y=\$2) is the ΔG vs. mfpt plot, where the separation from the target basin is measured by a distance in time units.
ATTENTION: If pieces of trajectories from different simulations are concatenated, insert a line with the entry "0" to prevent them from being used as a continues timeseries. "0" must not be used otherwise.

Bibliography

[1] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics*, 70:1185–1195, 2008.

[2] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-dimensional barrier preserving free-energy projections of a beta-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B*, 112:8701–8714, 2008.

[3] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA.*, 101:14766–14770, 2004.

[4] S. V. Krivov and M. Karplus. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B*, 110:12689–12698, 2006.

[5] M. Apaydin, D. Brutlag, C. Guesttin, D. Hsu, and J. Latombe. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. *Bioinformatics*, 18:18–26, 2002.